PREDICTION OF IMPACTS AND OUTBREAK OF COVID-19 USING DISTINCT LEARNING ALGORITHM

TASPIA TAZRI CHAITY

BACHELOR OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

PREDICTION OF IMPACTS AND OUTBREAK OF COVID-19 USING DISTINCT LEARNING ALGORITHM

TASPIA TAZRI CHAITY

This report was submitted in fulfillment of the requirements for the award of the degree of Bachelor of Science in Information and Communication Engineering

Department of Information and Communication Engineering

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

MARCH 2023

DECLARATION

This thesis is submitted to the Department of Information and Communication Engineering ofNoakhali Science & Technology University in partial fulfillment of the requirements for the award of the degree of Bachelor of Science (B.Sc) Engineering in Information and Communication Engineering (ICE). So, I hereby declare that this report has not been submittedelsewhere for the requirement of any kind of degree, diploma, or publication.

Taspia Tazri Chaity
ID: BKH1811013F
Session: 2017-18
Year: 4, Term: II
Department of Information and Communication Engineering
Noakhali Science and Technology University
Noakhali- 3814

ACCEPTANCE

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science (B.Sc) degree inInformation & Communication Engineering.

Professor Dr. Md. Ashikur Rahman Khan
Chairman
Department of Information and Communication Engineering
Noakhali Science and Technology University
Noakhali- 3814.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Dr. Md Ashikur Rahman Khan, for his immense support and guidance throughout my bachelor's degree here at Noakhali Science & TechnologyUniversity. Without his assistance, this thesis paper would be extremely difficult for me to complete. I am quite appreciative of all of his assistance. His pragmatism and advice helped me throughout my journey here. I am sure that the lessons learned will help me in my future endeavors.

My sincere thanks to all my classmates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways. And also thanks to the authorities of those academic institutions who approved us to collect data. Last but not the least, I would like to thank my mom and dad for encouraging me and standing throughout my life - words don't do justice to everything you've done for me.

ABSTRACT

The COVID-19 pandemic has caused significant negative impacts on daily wage workers in various professions due to restrictions such as lockdowns and social distancing measures. To analyze and predict these impacts, we collected data on 1665 daily wage workers from eight different professions through an offline survey. After preprocessing the data to remove duplicates, null values, and cleaning, we encoded categorical features and split the dataset into training and testing sets. We used seven machine learning algorithms including logistic regression, decision tree, random forest, support vector machine, AdaBoost, KNN, and XGBoost to develop a predictive model. Our study found that the random forest algorithm had the highest accuracy of 92.1%, followed by XGBoost with 91% accuracy. Our research highlights the potential of machine learning techniques in predicting the impacts of COVID-19 on daily wage workers in various professions. Our thesis paper serves as a guide for policymakers and government officials to understand the effects of COVID-19 on daily wage workers and to implement measures to mitigate these negative effects. Furthermore, the predictive model developed in this study can help predict the impact of future pandemics or crises on these workers and support the development of targeted policies and interventions to help them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND STUDY

The COVID-19 pandemic has affected the lives of people worldwide, causing not only a health crisis but also an economic crisis. Daily wage workers, who rely on their daily earnings to meet their basic needs, have been significantly impacted by this crisis. These workers are often employed in sectors that require physical presence and interaction with customers, making them more vulnerable to the virus's spread. According to the International Labor Organization (ILO), the COVID-19 pandemic has caused a significant decline in working hours and earnings for workers worldwide, with the most significant impact on the informal economy. In India, for example, the informal sector employs over 90% of the workforce, including daily wage workers in various professions such as CNG drivers, auto-rickshaw drivers, tailors, plumbers, retailers, construction workers, rickshaw drivers, and carpenters. The lockdowns and social distancing measures implemented to control the spread of the virus have resulted in a sharp decline in demand for their services, leading to loss of employment and reduced income. The pandemic's impact on these workers' lives goes beyond economic challenges, with many also facing health risks as they continue to work during the pandemic. Workers in sectors such as transportation, retail, and construction are at higher risk of contracting the virus due to the nature of their work. Moreover, these workers may lack access to adequate healthcare and may not have the financial resources to pay for medical expenses. To support daily wage workers during this difficult time, governments and other organizations have implemented various measures, such as providing financial assistance, food and essential supplies, and health and safety measures. However, the effectiveness of these measures in addressing the challenges faced by daily wage workers remains unclear, and further research is necessary to understand their impact on the ground. Therefore, this research paper aims to measure the impact of COVID-19 on the lives of daily wage workers in various professions and to analyze the measures taken by the government and other organizations to support them during this crisis. By providing a comprehensive understanding of the challenges faced by these workers, this study can inform policymakers and organizations in developing effective strategies to support them during and after the pandemic.

1.1.1 Machine Learning

Machine learning is a subfield of AI that enables machines to acquire knowledge and skills to determine their current capabilities. Machine learning is widely used in data mining, NLP, computer vision, speech recognition, embedded systems, modern robotics, biological data analysis, disease diagnosis, handwriting recognition, search engines, various fraud and spam detection, etc. used [1]. Machine learning is the process of learning from data. Analyzing millions of columns of data and manually building predictive models is impossible for humans. You can intelligently use your existing machines to perform a wide range of calculations. Machine learning is a more powerful technique than traditional data analysis and statistical tools. You can find hidden patterns in your data, learn from existing data, build predictive models, and apply them to new or similar data that is revealed. Machine learning could be a valuable tool for predicting the impact of COVID-19. By analyzing large datasets, machine learning algorithms can recognize patterns and predict future trends. Predicting impacts using machine learning techniques is a supervised learning task. By training a supervised learning algorithm on a dataset, we can predict the impact of COVID-19 on daily workers' income and financial stability with a certain level of accuracy.

1.1.2 Machine Learning Types

Supervised Learning:

A classifier learns on input data as examples and learns on related output variables as targets. It works very impressively for classification models. Output labels can be numeric or string values such as tags or classes. To identify the exact response when exposed to new data instances. Humans can supervise other people, so they can be considered one and the same. for example: When a teacher gives a student a good example to learn from, the student creates a regular rule from the given example. Some models belong to the classification category, some belong to the regression category [2].

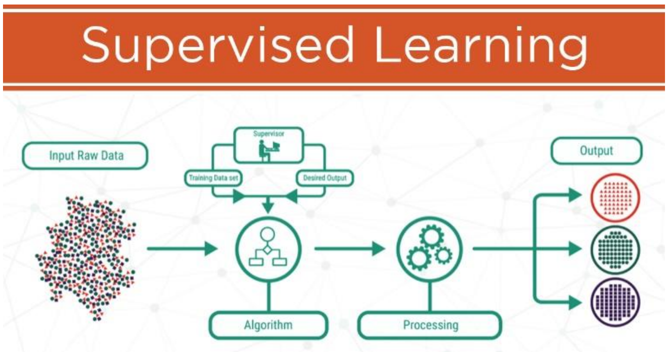


Figure 1. 1: Supervised Machine Learning Model

Unsupervised Learning:

Unsupervised learning is very different from supervised learning, which has no output variables as a supervised process. We do not find or predict anything specific here. Here we do what is called clustering or find patterns to group certain data together. The main difference with supervised learning is that we don't always need an output variable here [3]. A comparison between supervised and unsupervised Learning is shown in Figure

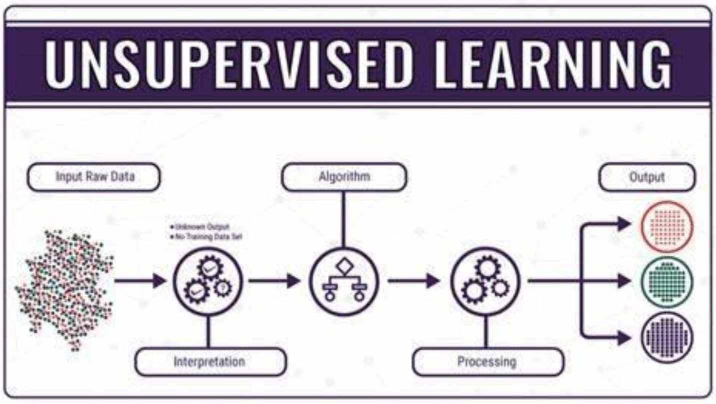


Figure 1. 2: Unsupervised Learning

Reinforcement learning:

This is called an ML technique that deals with executing actions from software agents. I'm leaning toward the DL method. It is one of the three main categories of machine learning. It's quite different from what we've done so far. Here we use the technique of assigning a prize to each good output and making them the basis of a classifier [4]. It can be trained and even used in neural networks. RL uses two specific learning models. they are:

Things like Markov decision processes are Q-learning.

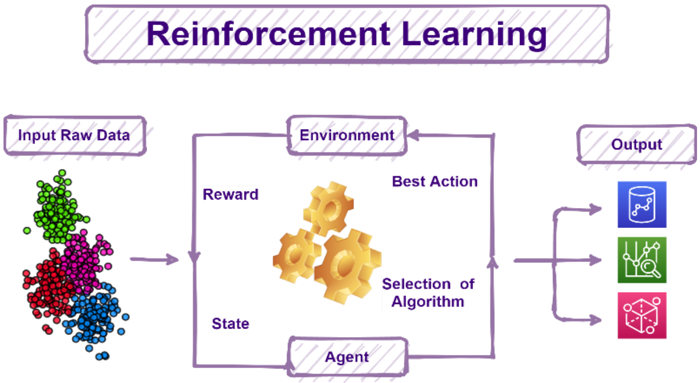


Figure 1. 3: Reinforcement learning

1.2 PROBLEM STATEMENT

The COVID-19 pandemic has caused a significant impact on the lives of daily wage workers worldwide. Daily wage workers, including CNG drivers, auto-rickshaw drivers, tailors, plumbers, retailers, construction workers, rickshaw drivers, and carpenters, are among the most vulnerable groups affected by the pandemic. These workers rely on their daily earnings to meet their basic needs, and the pandemic has caused a significant decline in demand for their services, leading to loss of employment and reduced income. Additionally, workers in these professions are at a higher risk of contracting the virus due to the nature of their work, further exacerbating their financial and health challenges. Governments and other organizations have implemented various measures to support these workers during this difficult time, such as financial assistance, food and essential supplies, and health and safety measures. However, the effectiveness of these measures in addressing the challenges faced by daily wage workers remains unclear. Therefore, the problem addressed in this research paper is to measure the impact of COVID-19 on the lives of daily wage workers in various professions and to analyze the measures taken by the government and other organizations to support them during this crisis. This study aim to provide a comprehensive understanding of the challenges faced by these workers, including the loss of employment, reduced income, and health risks, and the effectiveness of the measures implemented to support them. The findings of this research will inform policymakers and organizations in developing effective strategies to support daily wage workers during and after the pandemic.

1.3 OBJECTIVES OF THE STUDY

1. To analyze the impact of the COVID-19 pandemic on the daily income of daily wage workers in various professions.
2. To analyze the health conditions of daily wage workers during and after the pandemic.
3. To understand the difficulties faced by daily wage workers during the pandemic, such as loss of employment, reduced income, and health risks.
4. To preprocess and clean the collected raw data to ensure its accuracy and reliability for analysis.
5. To analysis the result using Machine learning algorithm.

1.4 SCOPE OF THE STUDY

This study involved the collection of pertinent data on daily workers, encompassing 1665 individuals from various professions such as CNG drivers, auto rickshaw drivers, rickshaw drivers, plumbers, construction workers, carpenters, tailors, and retailers. The number of data samples collected from each category was meticulously documented, with 262 from CNG Drivers, 220 from Auto Rickshaw Drivers, 151 from Rickshaw Drivers, 197 from Construction Workers, 232 from Tailors, 215 from Carpenters, 210 from Retailers, and 178 from Plumbers. Each data sample included 17 factors, including insightful personal information such as age, family size, number of working days per week, health conditions, and whether any family members were affected by COVID-19. Although most of the factors pertained to income rates, other variables were also taken into account, providing a comprehensive overview of the daily workers' experiences during the pandemic. After collecting data they are analyzed with machine learning algorithms. Thefollowing hardware, software requirements, programming language, and algorithm are used for analysis.

Hardware Requirement: Computer (Intel Core i3, Windows 10 operating system)

Software Requirement: Anaconda Navigator (version 1.7.2), Jupyter notebook.

Programming Language: Python (version 3.8.5).

Algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-NN, AdaBoost, XGBoost.

1.5 OUTLINE

In the following sections, we will present a comprehensive overview of our research. Chapter 2 provides a review of the relevant literature, while Chapter 3 outlines our methodology. In Chapter 4, we analyze the results of our study, and finally, Chapter 5 presents our conclusions.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

The COVID-19 pandemic has had a significant impact on societies worldwide, with various sectors and populations experiencing adverse effects. The literature review conducted thus far has explored the impact of the pandemic on diverse areas, including mental health, economy, agriculture, public transport, environment, and tourism.The study will utilize raw data collected through offline surveys to analyze the impact of the pandemic on daily workers. Furthermore, various machine learning algorithms will be applied to predict the potential impact of COVID-19 on this vulnerable population.

2.2 PREVIOUS WORKS

He et al. (2021) [5] conducted a study on employment changes of Chinese migrant workers in the post- COVID-19 era. They examined the impact of the pandemic on the employment status and income of Chinese migrant workers and their coping strategies. The study found that COVID-19 is having a significant impact on the employment status and earnings of migrant workers, with a higher proportion of workers experiencing job loss and reduced earnings. The authors recommend that the Chinese government provide more support to migrant workers during this difficult time. This study is relevant to ours because it highlights the impact of COVID-19 on income and employment of day laborers. However, this study was limited to Chinese migrant workers only and did not include other types of day laborers, which may limit the possibility of generalizing the results to other situations.

Hossein et al. (2021) [6] examines the impact of the COVID-19 pandemic on the labor market in Bangladesh using both primary and secondary data sources. The study found that the pandemic has caused significant disruptions to the labor market, resulting in unemployment, short hours and low wages for many workers. The authors suggest that policy makers need to focus on developing policies to support affected workers, especially those in the informal sector who are particularly vulnerable to the economic shocks caused by the pandemic.

Maljaars (2023) [7] conducted an online poll with 196 autistic and 228 non-autistic adults from Belgium, the Netherlands, and the United Kingdom in order to learn more about how the COVID-19 pandemic affected both autistic and non-autistic adults. The research discovered that participants with autism had a variety of experiences in all facets of life. Adults with autism reported less negative impact on their social lives than adults without autism, but more negative impact on their health and access to support services. The research also highlighted the variety of coping mechanisms employed by adults with autism to lessen the stress brought on by shifting and ambiguous policies. Despite the study's limitations, such as the sample's lack of racial and cultural variety, the results point to the need for improved support for autistic people.

Hossain, M.M., Asadullah, et al. (2021) [8] wanted this article to examine domestic violence against women and children during the COVID-19 pandemic. They collected data from 511 families on domestic violence against women and children during the COVID-19 outbreak. Participants received questionnaires to complete over 10 days. For this purpose, they used a machine learning based model. We applied random forest, logistic regression, and naive Bayesian machine learning algorithms to our model to predict domestic violence from our dataset. They used an oversampling strategy called the Synthetic Minority Oversampling Technique (SMOTE) for data balancing to improve model performance, used the statistical chi-square test to solve the imbalance problem, and analyzed the features of the dataset. determined the importance of the results of this study show that domestic violence increased in Bangladesh during the COVID-19 pandemic. In this paper, we monitored model effectiveness for three ML algorithms (imbalanced and balanced datasets). Experimental results for imbalanced data show that the models for the RF, LR, and NB algorithms have 64%, 63%, and 58% domestic violence prediction accuracies, respectively. For balanced data, the predictive accuracies of domestic violence for RF, LR, and NB classifiers are 77%, 69%, and 62%, respectively.

Štreimikienė, D. et al. (2022) [9] By treating Covid-19 like a global catastrophe and applying a vulnerability and resilience approach, this paper aimed to analyze the negative impact of Covid-19 on agriculture and food systems. The main contribution of this paper is to identify the most important measures to improve the resilience of agricultural systems in the face of the Covid-19 pandemic, based on recent scientific research, economics and measures published in 2020. It's about organizing and grouping. Current research is limited as it aims to capture only the general and rapid impact of the Covid-19 disaster on agricultural systems in the broadest sense. Future research is needed to examine how disruptions in various food supply markets affect small- and large-scale farms. Adverse impacts on food security need further investigation to define changes needed to make food systems more resilient

Banna, M. H. A. et al. (2022) [10] aimed to explore the impact of the COVID-19 pandemic on mental health among a representative sample of home-quarantined Bangladeshi adults. A cross-sectional design was used with an online survey completed by a convenience sample recruited via social media. The survey was split into three sections: participant characteristics, perceptions regarding COVID-19, and mental health. A total of 1,427 respondents were recruited. The prevalence of anxiety symptoms and depressive symptoms was 33.7% and 57.9%, respectively, and 59.7% reported mild to extremely severe levels of stress. The prevalence of mental health conditions in the adult population was lower (6.5% −31.0%) in pre-COVID-19 Bangladesh, which suggests that the pandemic may be responsible for increases in impaired mental health.

Kumar, S. et al. (2020) [11] provides a brief background on the COVID-19 outbreak and examines the impact of this pandemic on Bangladesh's tourism industry. The spread of COVID-19 is expected to have a long-term negative impact on tourism in Bangladesh and around the world. A secondary research approach was chosen for this study. The study looked at and analyzed various articles, newspapers and data from the World Tourism Organization (WTO) and World Health Organization (WHO). IATA predicts that global aviation revenues will decline by 11% in 2020, resulting in a loss of US$163 billion (IATA, 2020). The United Nations World Tourism Organization estimates that global international tourist arrivals could fall by 58% to 78% in 2020, resulting in a loss of $0.9

trillion to $1.2 trillion in international tourism receipts. As of June 30, 2020, over 2 million flights have been cancelled. About 65.5 million jobs are linked to the aviation sector, and IATA (2020) predicts that about 25 million aviation-related jobs are at risk worldwide. Planned trips have dropped by 80-90% in many cities around the world, such as museums, amusement parks, and sports venues closed down. The United Nations Conference on Trade and Development released a report in June 2021 stating that the global economy could lose over US$4 trillion as a result of the pandemic.

In this paper Cheval et al.(2020) [12] aimed to represents an early evaluate of the discovered and capacity affects of the COVID-19 at the environment. The predominant methodologic task of this evaluate changed into to acquire in a totally brief time applicable proof associated with the environmental effect of COVID-19 pandemic. Primary literature from the principle clinical flux changed into tested the usage of Thomson Reuters—Web of Science, Scopus and Google Academic. Press releases of global organizations, along with the World Health Organization and the World Meteorological Organization, had been precious reassets for legitimate records. In this paper they stated that, it really well worth citing the abundance of record of information to be had in a in particular brief time, which demonstrates the large hobby for the topic. For this reason of this investigation, they cite records from 142 bibliographic references (i.e.,fifty four blogs, 2 books, 2 chapters, seventy three peer-evaluation articles and eleven reports), and 116 websites, however they admire that they consulted at the least a double wide variety of reassets. The gathered proof includes information and records, observations, viewpoints and research posted at some point of the improvement of the COVID-19 pandemic, and that they had been decided on for his or her clean emphasis on environmental issues, city and public fitness matters, as argued within side the subsequent sections of this paper.

A research was done by McCartney et al. (2022 [13] to determine how COVID-19 affected the intentions of hospitality and retail employees to leave their jobs. The research used a quantitative methodology and used an online survey to gather data from 457 hospitality retail employees in the United Kingdom. Questions about work satisfaction, job stress, perceived organizational support, and perceived job security were included in the survey. Regression analysis was used in the research to look at the connection between turnover intentions and COVID-19. The COVID-19 pandemic significantly affected the intentions of hospitality and retail workers to leave their jobs, according to the study's findings. Particularly, employees with greater levels of job stress and lower perceptions of organizational support were more likely to plan to leave their jobs sooner. Additionally, the research discovered that employees who believed COVID-19 threatened their employment intentions were more likely to leave. The influence of particular COVID-19 policies or measures, which may have varying effects on turnover intentions, was not examined in the research.

In order to determine how the COVID-19 pandemic would affect Bangladesh's food security and vegetable supply network, Alam and Khatun (2021) [14] carried out a study. 304 farmers and 68 traders were surveyed as part of the study's mixed-method approach, and important informants from various points along the supply chain were also interviewed. The findings demonstrated that the pandemic had a significant negative effect on Bangladesh's vegetable supply chain, resulting in decreased output, transportation disruptions, and restricted access to markets. Farmers and traders alike experienced income losses and food insecurity as a consequence. The research also emphasized the significance of government actions in assisting the supply chain for vegetables and ensuring food security during emergencies like the COVID-19 pandemic.

A rapid review was conducted by Mhango et al. (2020) [15] to determine the risk factors for COVID-19 infection among healthcare professionals. The most frequent risk factors for COVID-19 infection among healthcare workers, according to a review of 13 articles by the authors, were a lack of personal protective equipment, prolonged contact with infected patients, and ignorance of COVID-19 prevention strategies. The review also emphasized how crucial it is to give health workers the right training, sufficient personal protective equipment, and regular screening. Due to the small number of included studies and the study's uneven geographic representation, there were some limitations. The results point to the need for practical measures to safeguard medical personnel and control the spread of COVID-19 in healthcare facilities.

Shen et al. (2020) [16] looked into how the COVID-19 pandemic affected business performance. From January to May 2020, the study gathered information from Chinese companies listed on the Shanghai and Shenzhen stock exchanges. According to stock returns and financial ratios, the pandemic had a detrimental effect on firm performance. Businesses in sectors like transportation, lodging, and catering as well as those with lower levels of profitability and liquidity were more negatively affected. However, the study also discovered that companies that performed better in terms of corporate social responsibility (CSR) suffered less from the pandemic. According to the study, companies need to improve their CSR practices in order to better withstand the damaging effects of pandemics and other crises.

The study by Kabir, Maple (2021) [17] investigated the impact of COVID-19 on Bangladeshi readymade garment (RMG) workers through an online survey. The sample included 579 RMG workers who were recruited through convenience and snowball sampling methods. The survey collected data on the workers' demographics, employment status, income, living conditions, and COVID-19 related information. The results showed that the pandemic had a significant impact on the workers' income, employment, and mental health. Many workers lost their jobs or experienced reduced working hours and wages, leading to financial hardship. The study had limitations in terms of a small sample size, non-random sampling, and reliance on self-reported data. Overall, the study suggests the need for policy interventions to support the RMG sector and improve the working conditions and well-being of its workers during and beyond the pandemic.

The COVID-19 pandemic's effects on sustainable production and operations management (SPOM) were studied by Kumar et al.(2020) [18]. The authors emphasized the impact of the pandemic on SPOM in terms of supply chain disruptions, consumer behavior changes, and the demand for adaptable and durable systems. The use of technology and innovation to address these issues and improve the sustainability of production and operations management was also covered. The paper offers a thorough analysis of the pandemic's effects on SPOM and possible long-term ramifications for sustainability.Data were gathered by the authors using a qualitative research methodology from a variety of sources, including reports, academic literature, and news articles. To find important themes and patterns, they used thematic analysis to analyze the data.

The paper by Masonbrink and Hurley (2020) [19] investigates how COVID-19 school closures affect kids and makes the case for child advocacy during this time. The authors highlight the potential drawbacks of school closures, such as disrupted learning and decreased access to vital services like nutrition and mental health support, by drawing on existing literature and research. They also talk about how important it is for decision-makers in the fields of healthcare, education, and policy to put children's needs first during a pandemic and prioritize their wellbeing. Based on a review of pertinent literature and policy documents, as well as their own professional experience working with children and families, the authors use a qualitative methodology.

This study by Sher, et al. [20] examines how the COVID-19 pandemic might affect suicide rates. Analysis of data from various nations and sources was done by the author as part of a review of the literature that has already been written on the subject. The paper emphasizes that the risk of suicide may be increased by the pandemic and its stressors, such as loneliness in social settings, financial strain, and infection fear. In addition, the author makes the case that some demographics, such as healthcare workers and people with mental health conditions already present, may be especially vulnerable. To fully understand and address this problem, more research is required, according to the paper, which also points out that the connection between the pandemic and suicide rates is complex and multifaceted. A review of the body of research on the topic, including studies and reports from different nations and sources, was part of the paper's methodology.

With the aid of machine learning techniques, Ardabili, S. F.et al. (2020) [21] sought to predict COVID-19 outbreaks. In order to predict the future spread of COVID-19, the authors gathered information on confirmed cases, deaths, and recoveries from a number of different countries and territories. The findings of the study demonstrated that future COVID-19 outbreaks could be accurately predicted using machine learning algorithms, particularly the Random Forest model. The study came to the conclusion that machine learning could offer precise and timely predictions of COVID-19 outbreaks and assist public health officials in making deft choices.

Table 2. 1: Summary of the literature review.

Citation	Paper Title	Technique Used	Parameter Used	Solution
K. Liu et al. [5]	Using mobile phone big data to discover the spatial patterns of rural migrant workers' return to work in China's three urban agglomerations in the post-COVID-19 era	Mobile phone big data analysis	Spatial patterns of rural migrant workers' return to work	Identification of key factors influencing migrant workers' return to work
N. Barker et al.[6]	Migration and the labour market impacts of COVID-19	Literature review	Labour market impacts	Analysis of the impact of COVID-19 on different types of migration and labour markets
J. Maljaars et al. [7]	Impact of the COVID-19 Pandemic on Daily Life: Diverse Experiences for Autistic Adults	Online survey	Impact of COVID-19 on daily life for autistic adults	Identification of specific challenges faced by autistic adults during the pandemic
F. Research et al. [8]	Violence Against Women and Children During COVID-19-One Year On and 100 Papers In	Literature review	Violence against women and children during COVID-19	Analysis of the prevalence and forms of violence during the pandemic and policy recommendations
D. Štreimikienė et al. [9]	Negative effects of COVID-19 pandemic on agriculture: systematic literature review in the frameworks of vulnerability, resilience and risks involved	Systematic literature review	Negative effects of COVID-19 on agriculture	Analysis of the impact of the pandemic on agriculture and policy recommendations
Md. H. al Banna et al. [10]	The impact of the COVID-19 pandemic on the mental health of the adult population in Bangladesh: a nationwide cross-sectional study	Cross-sectional survey	Impact of COVID-19 on mental health	Identification of risk factors for poor mental health during the pandemic
S. K. Deb et ai. [11]	Impact of COVID-19 Pandemic on Tourism: Perceptions from Bangladesh	Online survey	Impact of COVID-19 on tourism	Identification of specific challenges faced by the tourism industry during the pandemic
S. Cheval et al.[12]	Observed and Potential Impacts of the COVID-19	Literature review	Impact of COVID-19 on	Analysis of the observed and potential impacts of

	Pandemic on the Environment		the environment	the pandemic on the environment
G. McCartney et al.[13]	COVID-19 impact on hospitality retail employees' turnover intentions	Survey	N/A	Identified factors influencing turnover intentions among hospitality retail employees during the pandemic
G. M. M. Alam et al.[14]	Impact of COVID-19 on vegetable supply chain and food security: Empirical evidence from Bangladesh	Survey	N/A	Examined the impact of the pandemic on the vegetable supply chain and food security in Bangladesh
M. Mhango et al. [15]	COVID-19 Risk Factors Among Health Workers: A Rapid Review	Literature Review	N/A	Identified risk factors for COVID-19 among health workers based on previous research
H. Shen et al. [16]	The Impact of the COVID-19 Pandemic on Firm Performance	Survey	N/A	Examined the impact of the pandemic on firm performance in China
H. Kabir et al.[17]	The impact of COVID-19 on Bangladeshi readymade garment (RMG) workers	Survey	N/A	Explored the impact of the pandemic on Bangladeshi RMG workers
A. Kumar et al.[18]	COVID-19 impact on sustainable production and operations management	Literature Review	N/A	Examined the impact of the pandemic on sustainable production and operations management

A. R. Masonbrink et al. [19]	Advocating for Children During the COVID-19 School Closures	Literature Review	N/A	Discussed the potential negative impact of COVID-19 school closures on children and advocated for measures to support them
L. Sher et al.[20]	The impact of the COVID-19 pandemic on suicide rates	Literature Review	N/A	Examined the potential impact of the pandemic on suicide rates
S. F. Ardabili et al. [21]	COVID-19 Outbreak Prediction with Machine Learning	Machine Learning	COVID-19 data	Developed a machine learning model to predict COVID-19 outbreak

2.3 SUMMERY

The previous literature review mainly focuses on the impacts of COVID-19 on various sectors such as mental health, economy, agriculture, public transport, environment, and tourism. In contrast, our proposed technique aims to provide a comprehensive overview of the impacts of COVID-19 on a specific profession, i.e., daily workers. This study will include different types of daily workers, such as those in construction, transportation, and service sectors, who have been severely affected by the pandemic. We will analyze the impact of COVID-19 on daily workers using raw data collected through offline surveys. Additionally, we will apply various machine learning algorithms to predict the potential impact of COVID-19 on this vulnerable population.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This section outlines the methodology used in conducting the study. It begins with an overview of the research process in section 2.2, followed by a description of the tools utilized in section 2.3. Next, the dataset parameters, data collection process, and visualization of some dataset features are described in detail. The data preprocessing techniques, including dataset encoding, up-sampling, standard scaling, and splitting, are explained in sections 2.4 through 2.6. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-NN, Ada Boost, and XGBoost classifiers are used to classify the impact of COVID-19 on daily workers. Each algorithm is introduced in the corresponding section. The performance of these models is evaluated using various metrics such as confusion metrics, correlation matrix, accuracy, precision, recall, etc. Section 2.6 describes these evaluation metrics in detail. The application of these metrics helps in assessing the data and improving the accuracy of the research results.

3.2 RESEACH DESIGNS

The design of the system as shown in Figure 2.1 follows a typical machine learning pipeline for predictive modeling. The first step in this process is data collection from daily wage workers, and the second step is data preprocessing. Data cleaning is also an important part of the data preprocessing step, which involves removing any unclear, duplicate, or missing values. Label encoding is a technique used in machine learning to convert categorical data into numerical form, which is often necessary for modeling purposes. Therefore, it is also a part of the data preprocessing step. Once the data is preprocessed, the next step is to design the models. Our proposed system involves designing supervised machine learning models to predict the impact of COVID-19 on Daily Workers and classify them into specific classes, using preprocessed data. We will use various machine learning algorithms for this experimentation. The final step is to evaluate the accuracy of these models using multiple performance measurement techniques, and selecting the best accuracy among them. Each component of our proposed system is further elaborated in the following subsections.

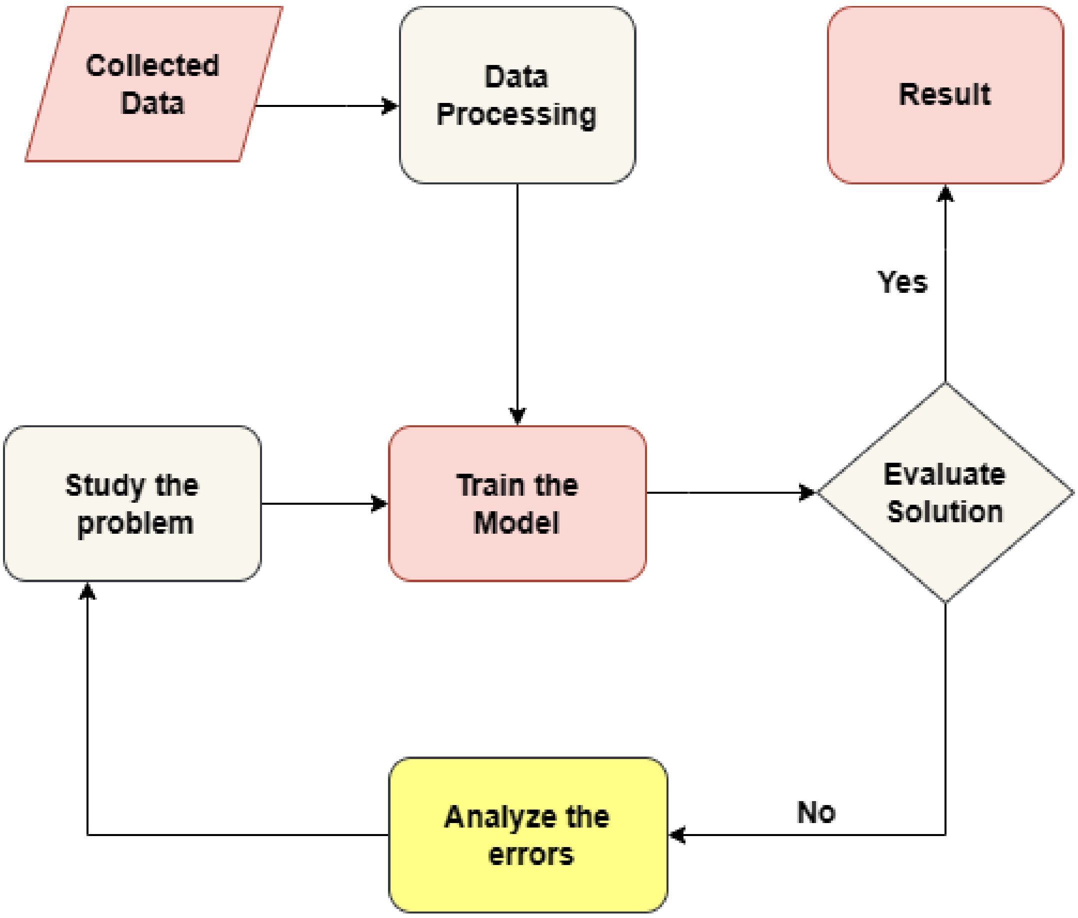


Figure 3. 1: Machine Learning System Overview

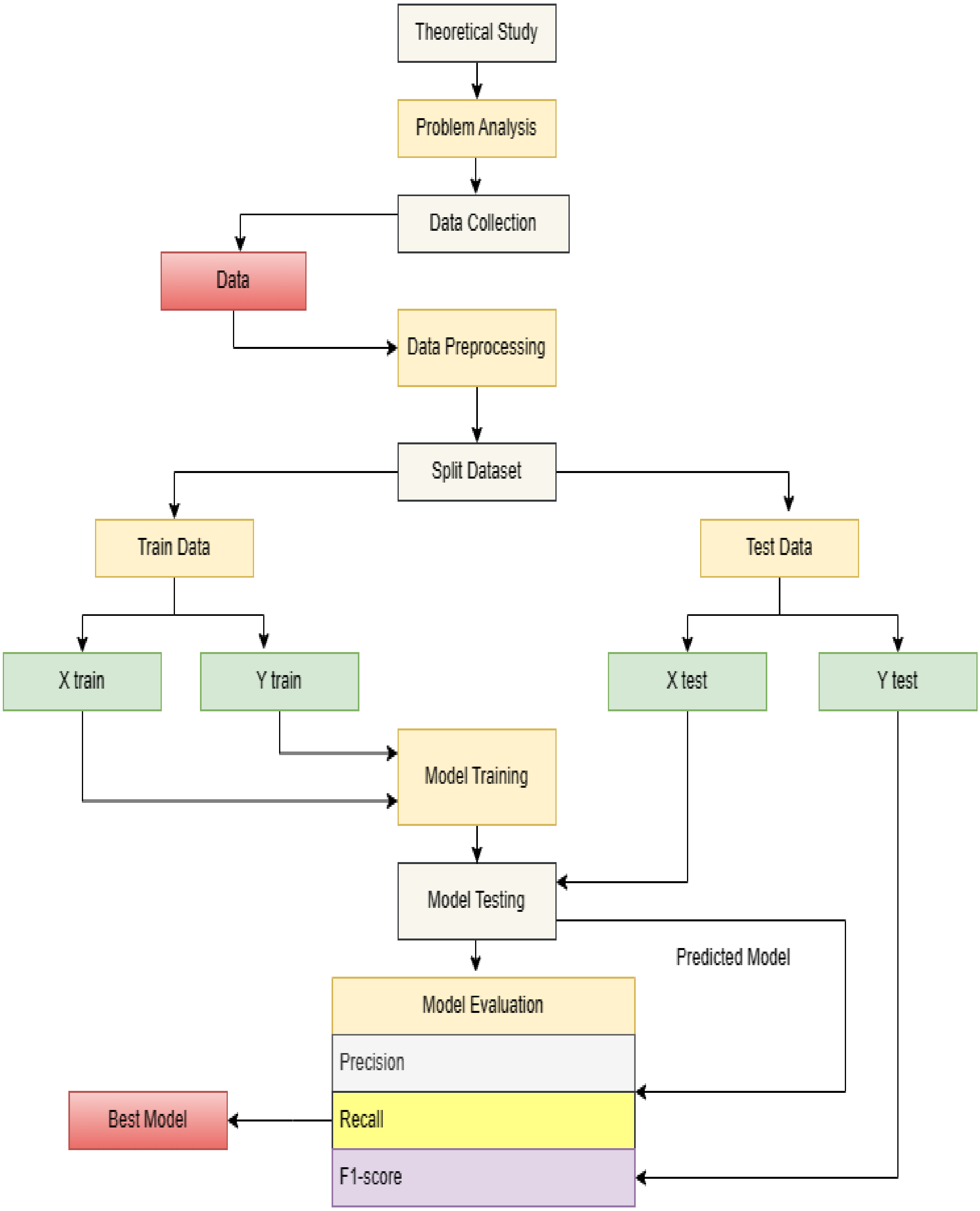


Figure 3. 2: Workflow Diagram

3.3 DATA COLLECTION

3.3.1 Parameters

To conduct research on predicting the difficulties faced by daily workers due to COVID-19, the following steps were taken:

1. Relevant data was collected from daily workers, comprising of 1665 workers' data from 8 types of daily workers including CNG Driver, Auto Rickshaw Driver, Rickshaw Driver, Tailor, Plumber, Retailer, Construction Worker, and Carpenter.
2. The data was collected through an offline survey by handing out questionnaires to the workers.
3. The survey considered about 17 factors, including not only their income information before, during, and after COVID-19, but also other personal information.

Table 3. 1: Feature Description

Feature Name	Description	Data Type
Name	Refers to the name of the worker	Object
Age	Age of the worker (in years)	Integer
Profession	Occupation of the worker	Object
Number of Members in Family	Count of individuals living in the worker's household	Integer
Income Before COVID-19 per day	Amount of money earned by the worker per day before the outbreak of COVID-19 pandemic	Integer
Income Before COVID-19 per month	Amount of money earned by the worker per month before the outbreak of COVID-19 pandemic	Integer

Income During COVID-19 per day	Amount of money earned by the worker per day during the outbreak of COVID-19 pandemic	Integer
Income During COVID-19 per month	Amount of money earned by the worker per month during the outbreak of COVID-19 pandemic	Integer
Income After COVID-19 per day	Amount of money earned by the worker per day after the outbreak of COVID-19 pandemic	Integer
Income After COVID-19 per month	Amount of money earned by the worker per month after the outbreak of COVID-19 pandemic	Integer
Expense per day	Amount of money spent by the worker on a typical day	Integer
Expense per month	Amount of money spent by the worker per month	Integer
Number of days worked in a week	Frequency of work done by the worker in a week	Integer
Physical health condition after COVID-19	Self-reported physical health condition experienced by the worker after COVID-19	Object
Any member of your family affected by COVID-19	Indicates if any member of the worker's family has been affected by COVID-19	Object
Any difficulties faced due to COVID-19	Difficulties faced by the worker due to COVID-19	Object

3.3.2 Data Collection

The required data for this research was collected through offline surveys by distributing questionnaires. The questionnaires included factors that were previously discussed in the subsection. A total of 1665 daily workers participated in the survey, including CNG Drivers, Auto Rickshaw Drivers, Rickshaw Drivers, Tailors, Carpenters, Construction

Workers, Retailers, and Plumbers. The number of data samples collected from each category is as follows: 262 from CNG Drivers, 220 from Auto Rickshaw Drivers, 151 from Rickshaw Drivers, 197 from Construction Workers, 232 from Tailors, 215 from Carpenters, 210 from Retailers, and 178 from Plumbers. Each data sample included 17 factors.

Table 3. 2: Daily Worker Types and Data Sample Sizes

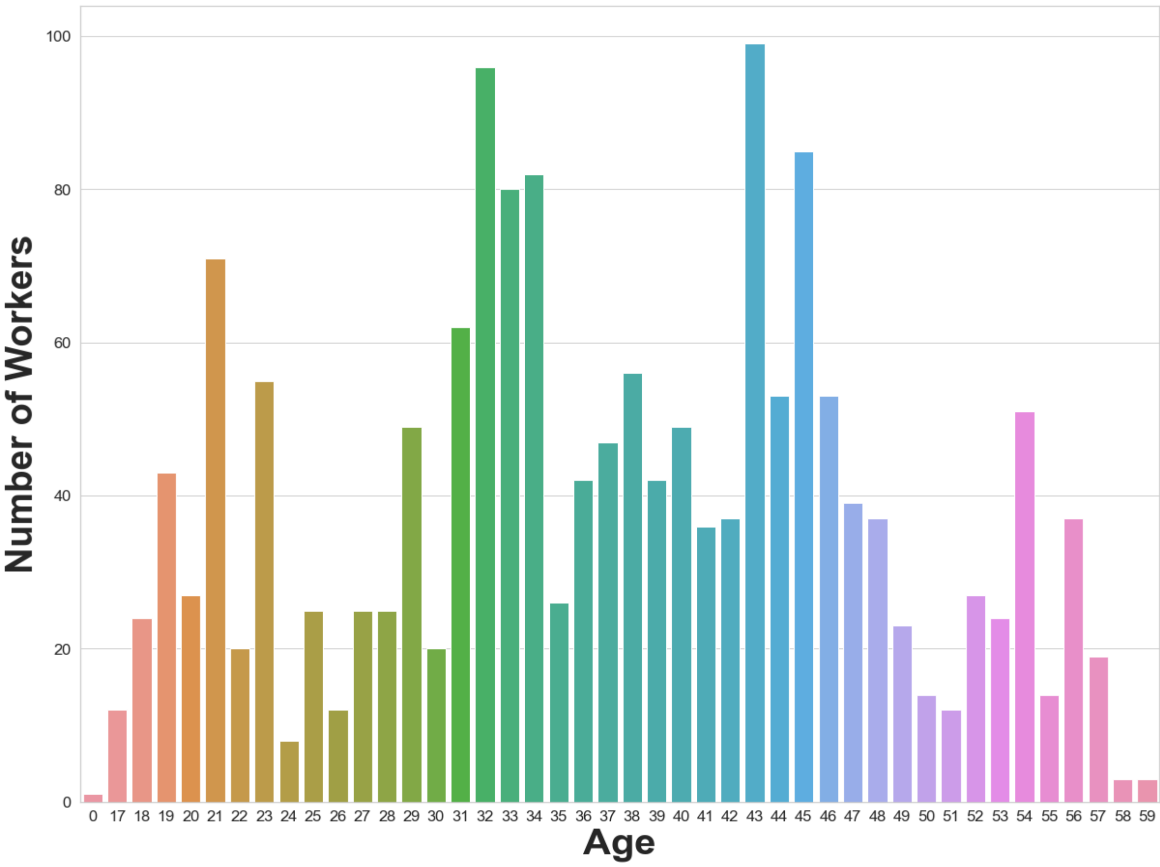
Daily Worker Type	Data Sample Size
CNG Driver	262
Auto Rickshaw Driver	220
Rickshaw Driver	151
Tailor	232
Construction Worker	197
Carpenter	215
Retailer	210
Plumber	178

3.3.3 Dataset Description

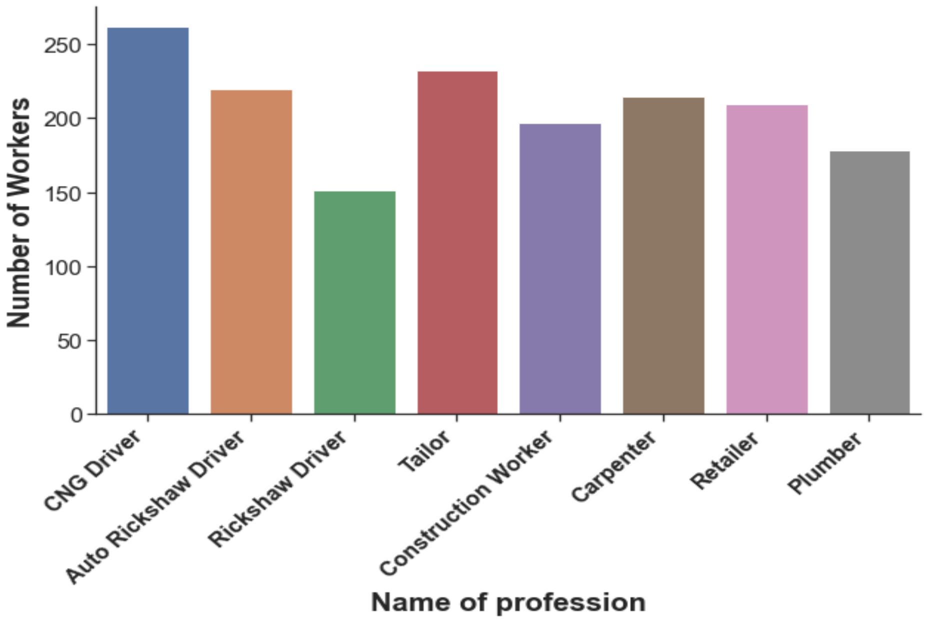
In this part, we examine the gathered information and give a description of the dataset.17 features were chosen for the dataset in order to effectively use the data. We're going to prepare and use this information to teach the models. So let's first examine our information in more detail.

Table 3. 3: Data Sample

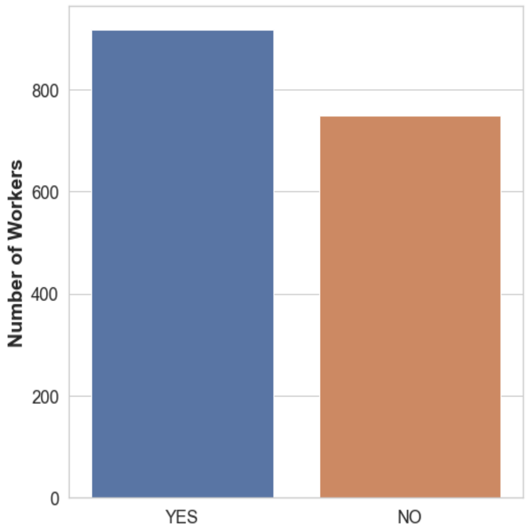
Sl no.	Name	Age	Profession	Number of members in family	Before COVID-19 Per day (TK.)	Before COVID-19 Per Month (TK.)	During COVID-19 Per day (TK.)	During COVID-19 Per Month (TK.)	After COVID-19 Per day (TK.)	After COVID-19 Per Month (TK.)	Expense Per day (TK.)	Expense Per Month (TK.)	Number of days work in a week	Physical health condition after COVID-19 as you feel	Any member of your family affected by COVID-19
1	Shariful Hossain	32	CNG Driver	5	900	19800	300	6600	600	13200	400	8800	5	Joint Pain	YES
2	Md. Monir Hossain	35	CNG Driver	6	1200	31200	600	15600	800	20800	450	11700	6	Autoimmune Disease	YES
3	Md. Sumon Mia	31	CNG Driver	4	1100	19800	600	10800	800	14400	500	9000	4	Muscle Pain	YES
4	Md. Sumon Mia	28	CNG Driver	7	1500	33000	800	17600	1600	35200	700	15400	5	Muscle Pain	NO
5	Md. Jalal Hossain	30	CNG Driver	5	1200	31200	600	15600	1000	26000	600	15600	6	Chest Pain	YES
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1661	Imran	33	Carpenter	6	1350	29700	650	14300	850	18700	650	14300	4	Fatigue	YES
1662	Naeem	29	Carpenter	5	1150	25300	450	9900	550	12100	450	9900	6	Chest Pain	YES
1663	Shuvo	36	Carpenter	6	1500	33000	800	17600	1000	22000	700	15400	5	Asthema	NO



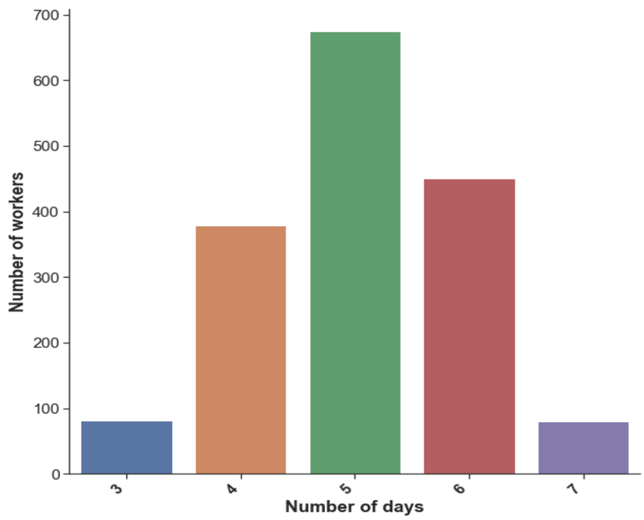
(a) Ages of collected sample data



(b) No. of workers versus profession



(c)Number of family members of the workers affected from COVID-19



(d) No. of workers and no. if days the workers work in a week

Figure 3. 3: Frequency distributions of the collected data.

3.4 DATA PREPROCESSING

This is one of the most important states of the implementation module. We may collect a large amount of data from several sources so some of the data might contain garbage values, some may be null. Mostly, there can be unclear, duplicate, and null values in the dataset. Also, some data may contain large dimensions that can be bothersome because it can significantly increase the training time. That’s why data preprocessing is a vital part of this process. Data preprocessing includes procedures to handle duplicates, null values, dimension reduction, etc. The following Figure 2.2 displayed below indicates the steps involved in data preprocessing.

3.4.1 Data Cleaning

Data cleaning is an important step in data preprocessing that involves removing or correcting any inconsistencies, errors, or irrelevant data from the dataset. This step helps to ensure that the data is accurate, complete, and consistent, which is essential for accurate analysis and modeling. One common data cleaning technique is removing duplicate values, which involves identifying and removing any rows or entries that are exact duplicates of one another. This can be done using the drop_duplicates() method in pandas. Another technique is removing any rows or entries with missing or NaN values, which can be done using the dropna() method in pandas. In addition, the drop() method can be used to remove specific rows or columns from the dataset based on certain conditions or labels. The iloc[] method can be used to select specific rows or columns from the dataset based on their indices or labels. Lastly, the replace() method can be used to replace any values in the dataset with a specified value, such as replacing any 0 values with a unique value to avoid any confusion or bias during analysis. Overall, data cleaning is an essential step in data preprocessing that involves identifying and correcting any issues in the dataset to ensure accurate analysis and modeling.

3.4.2 Encoding the Dataset

In machine learning, we frequently work with datasets that include numerous labels in one or more columns. Labels in the form of words or numbers can be used. The training data is frequently labeled in words to make it intelligible or human-readable. When it comes to digesting this type of data, machine learning algorithms aren't very good. We must first preprocess such data before feeding it to an algorithm. To put it another way, we need to apply some modifications to it. So, we must convert our data into a numerical form. These are called dummy variables. Scikit-Learn provides three very good encoder classes Label Encoder, One Hot Encoder, and Ordinal Encoder. We employed Scikit-Learn's LabelEncoder class to encode our dataset. Label encoding translates the labels into a numeric format so that they can be read by the machine learning algorithms. The following Table 2.5 shows the output of the label encoding process on the dataset.

Table 3. 4 After Label Encoding the dataset**.**

Sl no.	Name	Age	Profession	Number of members in family	Before COVID-19 Per day (TK.)	Before COVID-19 Per Month (TK.)	During COVID-19 Per day (TK.)	During COVID-19 Per Month (TK.)	After COVID-19 Per day (TK.)	After COVID-19 Per Month (TK.)	Expense Per day (TK.)	Expense Per Month (TK.)	Number of days work in a week	Physical health condition after COVID-19 as you feel	mem of y fa affe COVIL
438	16	48	4	7	1100	24200	400	8800	1400	30800	900	19800	5	1	
598	101	55	1	9	1200	21600	700	12600	1300	23400	600	10800	4	4	
1293	177	21	4	3	500	11000	200	4400	1000	22000	500	11000	5	3	
1287	219	42	4	5	600	10800	300	5400	700	12600	400	7200	4	6	
1208	96	33	5	5	400	8800	200	4400	500	11000	400	8800	5	5	
579	23	55	1	9	1200	26400	500	11000	1000	22000	400	8800	5	6	
1325	27	33	2	5	500	9000	300	5400	700	12600	500	9000	4	9	
860	419	54	4	6	1000	26000	500	13000	1100	28600	900	23400	6	1	
703	308	36	0	5	600	18000	300	9000	800	24000	500	15000	7	10	
757	200	37	1	5	400	8800	200	4400	500	11000	400	8800	5	5	
234	69	19	3	4	400	10400	300	7800	700	18200	300	7800	6	0	
358	5	53	5	7	700	9800	400	5600	600	8400	400	5600	3	10	
161	282	32	7	6	1000	26000	300	7800	1200	31200	500	13000	6	9	
361	31	43	5	5	1600	41600	700	18200	1200	31200	800	20800	6	5	
1233	123	45	7	4	300	5400	200	3600	600	10800	300	5400	4	0	
672	449	48	2	7	800	17600	300	6600	600	13200	500	11000	5	11	
1567	399	48	7	3	700	12600	400	7200	600	10800	300	5400	4	6	
1462	336	39	0	6	1000	21000	600	12600	800	16800	400	8400	5	9	
571	169	34	3	5	900	23400	300	7800	1200	31200	400	10400	6	10	
1027	40	45	1	5	800	17600	400	8800	1000	22000	700	15400	5	6	

3.4.3 Up sampling The Data

Up sampling is a data preprocessing step used to handle class imbalance, where the number of samples in one or more classes is significantly lower than the others. This can cause a machine learning model to be biased towards the majority class and lead to poor

performance. To address this issue, up sampling involves increasing the number of samples in the minority class(es) to balance the number of samples between classes.
In the code snippet provided, the resample() function from the scikit-learn library is used to up sample the minority classes. The function takes three parameters: replace, n samples, and random state. replace=True means that sampling is done with replacement, which allows the same sample to be selected multiple times. N samples specifies the number of samples to generate for each class, which is set to 4000 in this example. Finally, random state is used to ensure that the same samples are generated each time the code is run. The code separates the data into three classes based on the value of the 'Any difficulties face Due to COVID-19' column, namely the majority class with value 0 and the two minority classes with values 3 and 4. It then applies the resample() function to each minority class, generating 4000 samples for each class. The concat() function is then used to combine the up sampled minority classes with the original majority class, resulting in a new dataset with a balanced number of samples for each class. The technique used in this code is known as random up sampling, where new samples are generated by randomly selecting existing samples from the minority class(es) and duplicating them. There are other techniques that can be used for up sampling as well, such as SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples based on the patterns in the minority class.

3.4.4 Standard Scaling

In order to prepare the data for modeling, a standard data preprocessing technique known as Standard Scaling was employed. This technique involves transforming the data in a way that the mean of each feature is centered at 0 and the standard deviation is scaled to 1. The StandardScaler class from the sklearn.preprocessing library was used to perform this operation on the training and testing sets. First, the scaler was fit on the training set using the fit_transform() method, which computes the mean and standard deviation of each feature in the training set and applies the transformation to the training set.
Then, the same transformation was applied to the testing set using the transform() method, ensuring that the testing set is scaled in the same way as the training set. This is important to ensure that the model is not biased towards any particular subset of the data.

3.4.5 Splitting the Dataset

It is not recommended to use the entire dataset for training a model. Instead, it is customary to split the dataset into two separate sets: a training set and a testing set. The purpose of this separation is to evaluate the model's performance on unseen data. The majority of the dataset is used for training, typically around 80%, while the remaining portion is reserved for testing. The separation is usually done randomly to ensure a good representation of the data in both sets. By testing the model on data that it has not seen during training, the effects of data discrepancies can be minimized, and a better understanding of the model's characteristics can be obtained.The training set is used to train the model, and after that, the model is evaluated on the testing set by making predictions against it. Since the testing set contains known values for the attribute to be predicted, it is easy to assess whether the model's predictions are correct. Testing is the last step in the process, and it is used to check if the trained model generalizes well with the dataset. The Scikit-Learn train_test_splitclass is often used to split the dataset into training and testing sets.

3.5 MODEL DEVELOPMENT

This is the most important phase which includes building the models that will aid in the classification task. We implement four machine learning algorithms for prediction. These algorithms include Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, KNN, AdaBoost and XGBoost. All of these algorithms will be discussed in the following sub-sections.

3.5.1 Logistic Regression

Logistic regression is a popular machine learning algorithm that is primarily used for classification tasks. Its main objective is to estimate the likelihood that a given instance belongs to a specific class[22]. The algorithm calculates a probability score, and if the score exceeds 50%, the model predicts that the instance belongs to that class; otherwise, it predicts that it does not. Logistic regression is a binary classifier, meaning that it can only classify instances into two distinct categories. To predict the class of an instance, a threshold value is set, and the sigmoid function is typically used as the activation function for this algorithm. Logistic regression is especially useful for understanding how multiple independent variables influence a single outcome variable in classification problems. The model calculates a weighted sum of the input features, including a bias term, and then applies the logistic function to this sum before outputting the result.Overall, Logistic regression is a powerful tool for classification tasks and is widely used in various fields, including healthcare, finance, and marketing, among others.

$$\hat{p} = h_{\theta}(x) = \sigma(\theta^{T} . x) \tag{3.1}$$

Where $\sigma(.)$ is a sigmoid function that outputs a number between 0 and 1. It is defined as

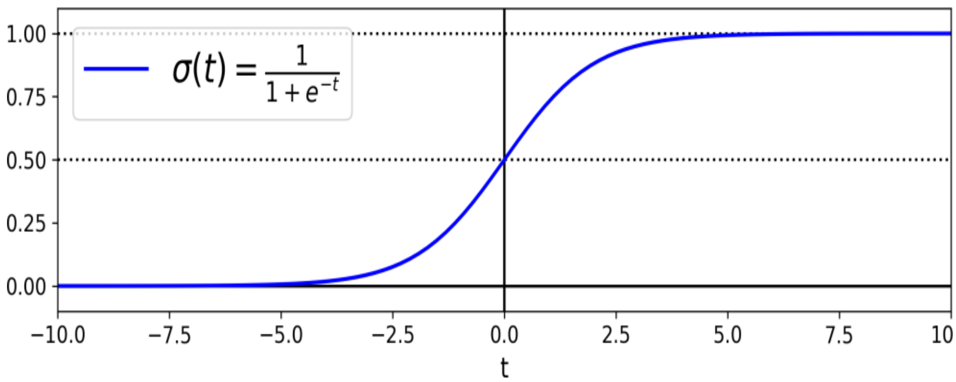


Figure 3. 4: Sigmoid function.

$$\sigma(t) = \frac{1}{1 + exp(-t)} \tag{3.2}$$

Once the Logistic Regression model has estimated the probability $\hat{p} = h_{\theta}(x)$ that an instance

x belongs to the positive class, it can make its prediction $\hat{y}$ easily.

$$\hat{y} = \{ 0 \text{ if } \hat{p} < 0.5, \tag{3.3}$$

$$1 \; if \; \hat{p} \geq 0.5.$$

Notice that $\sigma(t)$ < 0.5 when t < 0, and $\sigma(t) \geq 0.5$ when $t \geq 0$, so a Logistic Regression model predicts 1 if $\theta^T . \boldsymbol{x}$ is positive, and 0 if it is negative. Unfortunately, this algorithm worksonly when the predicted variable is binary, assumes all predictors are independent of each other,and assumes data is free of missing values. But it is possible to perform multi-classification tasks using the one-vs-all method. Scikit-Learn has the LogisticRegression class to implement this classification algorithm.

3.5.2 Random Forest

Random forest is a supervised learning algorithm. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method's basic premise is that combining several learning models improves the overall output. The Random Forest algorithm introduces extra randomness when splitting a node, it searches for the best feature among a random subset of features, this results in a greater tree diversity, which trades a higher bias for a lower variance, generally yielding an overall better model [23]. So the prerequisites for the random forest to perform well are: There needs to be some actual signal in our features so that models built using those features do better than random guessing. Also, the predictions (and therefore the errors) made by the individual treesneed to have low correlations with each other. Some of the features of this algorithm are statedbelow:

1. It takes less training time as compared to other algorithms.
2. It predicts output with excellent accuracy, and it runs quickly even with a huge dataset.
3. It can also maintain accuracy if a significant amount of data is absent.

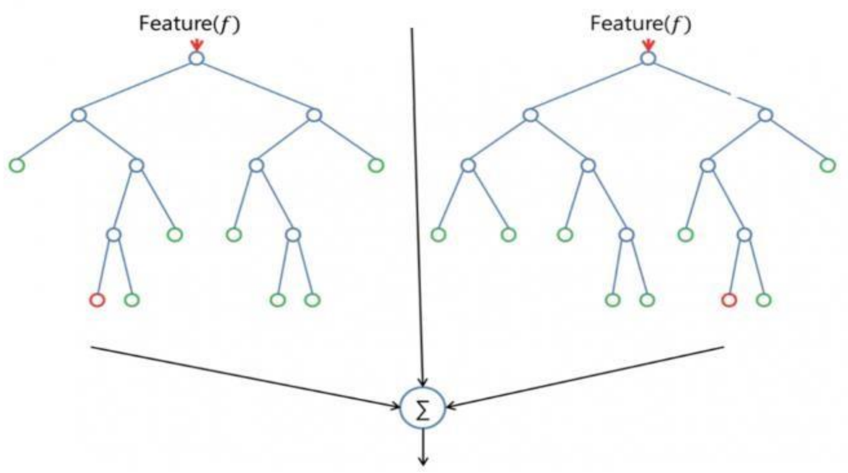


Figure 3. 5: Random Forest

The random forest is formed in two phases: the first is to combine the N decision trees to buildthe random forest, and the second is to make predictions for each tree created in the first phase.The following steps can be used to demonstrate the working process:

Step - 1. Pick K data points at random from the training set.

Step - 2. Create decision trees for the data points you've chosen (Subsets). **Step - 3.** Decide on the number N for the decision trees you wish to create.**Step - 4.** Repetition of Steps 1 and 2.

Step - 5. Find the forecasts of each decision tree for new data points, and allocate thenew data points to the category with the most votes.

Scikit-Learn has the RandomForestClassifier class to implement this classification algorithm.

3.5.3 Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions[24]. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Step 1. Start with the root node, which holds the entire dataset, say S.

Step 2. Using the Attribute Selection Measure, find the best attribute in the dataset

Step 3. Subdivide the S into subsets that contain the attribute's best possible values.

Step 4. Create the node of the decision tree that has the best attribute.

Step 5. Create additional decision trees in a recursive manner using the subsets of the dataset obtained in step 3. Continue this process until the nodes can no longer be classified, at which point the final node is referred to as a leaf node.

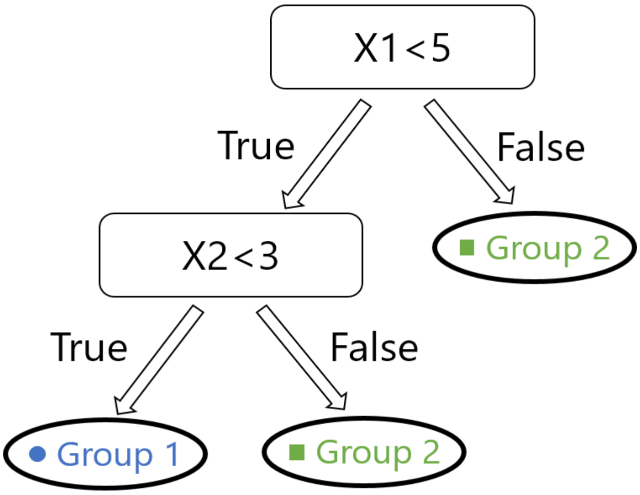


Figure 3. 6: Decision tree

The biggest challenge that emerges while developing a Decision tree is how to choose the bestattribute for the root node and sub-nodes. So, there is a technique called Attribute Selection Measure, or ASM, that can be used to overcome such situations. We can easily determine the best property for the tree's nodes using this measurement. The following are two popular ASMtechniques:

1.Information Gain - The assessment of changes in entropy after segmenting a dataset based on an attribute is known as information gain. It determines how much data a feature offers about a class. We split the node and build the decision tree based on the value of information gained. The highest information gain node/attribute is split first ina decision tree method, which always strives to maximize the value of information gain. The following formula can be used to compute it:

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j, after) \qquad (3.4)$$

2.Gini Index - The Gini index is a measure of impurity or purity used in the CART (Classification and Regression Tree) technique to create a decision tree. In comparisonto a high Gini index, an attribute with a low Gini index should be favored. It only makesbinary splits, and the CART method creates binary splits using the Gini index. The following formula can be used to compute the Gini index:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \qquad (3.5)$$

Scikit-Learn has the DecisionTreeClassifier class to implement this classification algorithm.

3.5.4 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane[25]. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

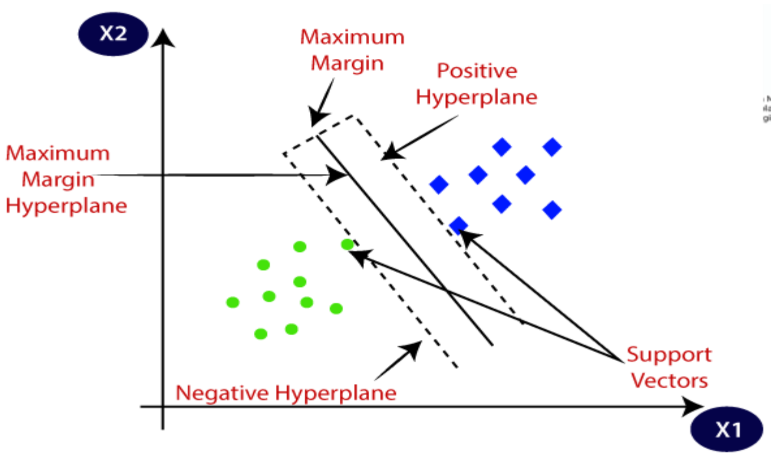


Figure 3. 7: Support Vector Machine

3.5.5 K- Nearest Neighbor Algorithm:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm[26]. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**

- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

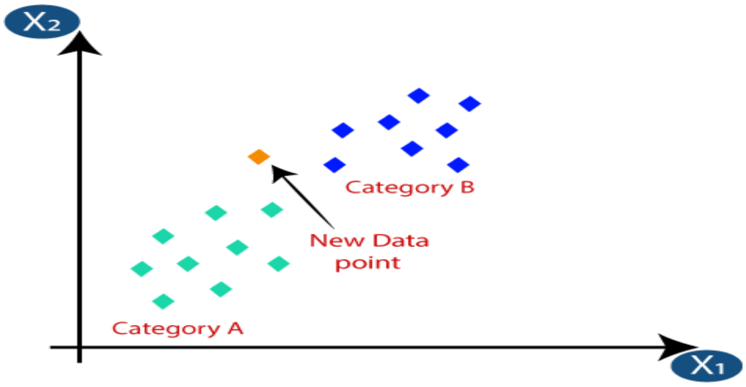


Figure 3. 8: K-NN

3.5.6 AdaBoost

AdaBoost is a classification meta-algorithm that adapts subsequent weak learners to favor instances misclassified by previous classifiers, making it less prone to overfitting than other learning algorithms. While individual learners can be weak, as long as each performs slightly better than random guessing, the final model will converge to a strong learner. AdaBoost is typically used to combine weak base learners, such as decision stumps, but can also effectively combine strong base learners, such as deep decision trees, for even greater accuracy. Every learning algorithm has strengths and weaknesses and requires tuning to achieve optimal performance on a given dataset. AdaBoost, particularly when used with decision tree learning, is often considered the best "out-of-the-box" classifier [27]. Information about the relative difficulty of each training sample is fed into the tree growing algorithm at each stage of the AdaBoost process, allowing later trees to focus on more challenging examples.

3.5.7 XGBoost:

XGBoost is a powerful machine learning library that uses optimized distributed gradient boosting to efficiently train models on large datasets. It is an ensemble learning method that combines multiple weak models to create a stronger prediction and has gained popularity due to its ability to achieve state-of-the-art performance in tasks like classification and regression. One of XGBoost's notable strengths is its efficient handling of missing data, eliminating the need for extensive pre-processing of real-world data[28]. Additionally, it supports parallel processing, enabling models to be trained on massive datasets in a reasonable timeframe. XGBoost is a versatile tool applicable to various domains, such as Kaggle competitions, recommendation systems, and click-through rate prediction, among others. Furthermore, it is highly customizable, allowing for fine-tuning of various model parameters to achieve optimal performance.

3.6 PERFORMANCE ANALYSIS

Performance analysis includes making methodical observations to improve performance, which helps with data decision-making and provides feedback for our system. It is the process of analyzing or assessing how well a specific situation performs in light of the goal to be attained. We will evaluate our results with various performance measures accuracy, confusion matrix, precision value, f1-score.

3.6.1 Confusion Matrix

A Confusion matrix is an N × N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. Confusion Matrix gives output as a matrix that demonstrates the total performance of a classification model or algorithm for machine learning processes.

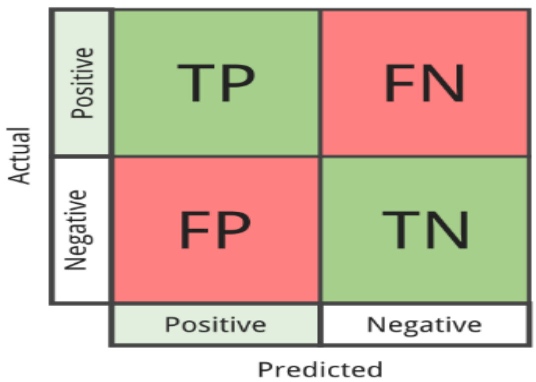


Figure 3. 9: The Confusion Matrix

The target variable in a binary classification has two values: Positive or Negative. The columns represent the predicted values of the target variable. The rows represent the actual values of the target variable.

1. **True Positive (TP):** The predicted value matches the actual value. The actual value was positive and the model predicted a positive value.
2. **True Negative (TN):** The predicted value matches the actual value. That means the actualvalue was negative and the model predicted a negative value
3. **False Positive (FP) – Type 1 error:** The predicted value was falsely predicted. That means the actual value was negative but the model predicted a positive value. Also knownas the Type 1 error.
4. **False Negative (FN) – Type 2 error**: The predicted value was falsely predicted**.** That means the actual value was positive but the model predicted a negative value**.** Also knownas the Type 2 error.

$$Accuracy = \frac{TN+FN}{N}$$, where(N= Total number of Samples) (3.6)

3.6.2 Performance Evaluation Metrics

The following performance evaluation metrics are utilized to analyze the performanceof the classification models

Accuracy:

Accuracy is a performance metric used in machine learning to measure how well a model correctly predicts the outcome of a binary or multiclass classification problem. It is the ratio between the total number of correct predictions to the total number of input sample. It is expressed as a percentage and ranges from 0 to 100%, with 100% indicating that the model made all correct predictions. It is a statistical method.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.7}$$

Precision:

Precision is a metric used to evaluate the performance of a classification model. It is the ratio of true positive predictions to the total number of positive predictions made by the model. Mathematically, precision can be expressed as:

$$Precision = \frac{TP}{(TP+FP)} \tag{3.8}$$

Recall:

It is the ratio of correct positive results and the total number of all associated samples. Equation of recall:

$$Recall = \frac{TP}{(TP+FN)} \tag{3.9}$$

F1 Score:

It is often convenient to combine precision and recall into a single metric called theF1 score We will use the F1 score to measure the test accuracy. We estimated F1 score balancing precision and recall. It is the harmonic mean of two (precision and recall). The value of f1 score ranges from o to 1.

Equation of f1 score:

$$F1\ score = \frac{2*precision*recall}{precision+recall} \tag{3.10}$$

3.7 FACTOR FOR ANALYSIS

The COVID-19 pandemic has had a significant impact on the world, affecting various aspects of life, including work and health. The pandemic has led to widespread job loss and financial instability, particularly among daily wage workers. The distinct problems faced by daily wage workers during the pandemic include a lack of social protection, poor working conditions, and inadequate access to healthcare. The impact of COVID-19 on the health condition of daily wage workers is another area of concern, with workers facing increased risks of contracting the virus due to their work environment and lack of access to protective measures. In addition, COVID-19 has had a significant impact on the income of daily wage workers, with many losing their jobs or facing reduced hours and wages. Finally, the pandemic has also affected different professions in distinct ways, with some professions facing more significant disruptions than others. Therefore, this paper we will examine the distinct problems faced by daily wage workers during the pandemic, the impact of COVID-19 on their health condition, income, and different professions, to provide insights into how policymakers can mitigate the effects of the pandemic on vulnerable workers.

3.8 COMPARISION OF PERFORMANCE METRICS

In this section of the study, we will compare the performance metrics of seven machine learning algorithms, namely Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, and XGBoost, that will be used for predicting the impact of COVID-19 on daily wage workers. The performance metrics that will be compared include accuracy, precision, recall, and F1 score.

3.9 FIND THE BEST MODEL FOR PREDICTION

In our model we have several machine learning model namely Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, and XGBoost to predict the impact of COVID-19 on daily wage workers. The model with the highest accuracy will be selected as the best one for for predicting the impact of COVID-19 on daily wage workers.

3.10 REQUIRED TOOLS

Different tools will be used for this study. All of them are free and open source. A short description of these tools is given below,

1. Python: Python is a high-level general programming language.
2. Virtual Environment: We will use Jupyter Notebook as the environment to conduct ourexperiment. Anaconda framework provides Jupyter Notebook along with other tools thatprovide assistance in conducting such experiments. We will create a virtual environmentinside Anaconda and install all the required tools.

3. Scikit-learn: Scikit-learn is a widely used popular library for machine learning.

4. NumPy: NumPy is a very powerful package that enables us for scientific computing.

5. Matplotlib: Matplotlib is a plotting library for Python and its numerical mathematics extension NumPy.

6. Pandas: Pandas is an open-source BSD licensed software specially written for python.

7. Seaborn: Seaborn is a library used for data visualization and is created by using python.

8. SciPy: SciPy is a collection of fundamental mathematical functions and is built on the top of Scikit-learn.

9. OS: The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system dependent functionality.

CHAPTER 4

RESULT AND DISCUSSION

4.1 INTRODUCTION

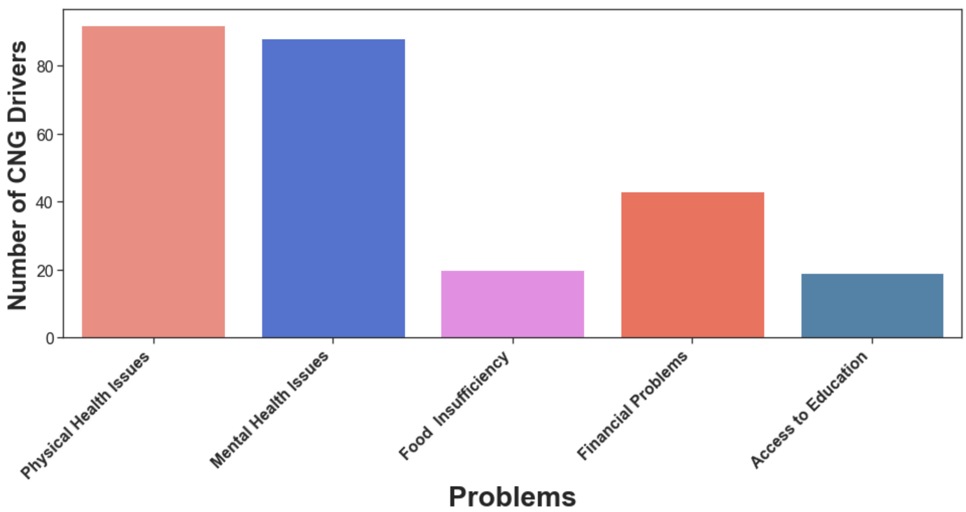
In this chapter, we evaluate the dataset by examining the results obtained from the exploratory analysis using Bar Charts, Correlation Matrix. This is included in the following section. Next, we discuss the results of the numerical analysis performed with Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-NN, Ada Boost and XGBoost. Finally, we'll conclude our discussion after comparing the results obtained by various performance metrics like Accuracy, Precision, Recall, and F1 score. Comparing these values will lead to the classification model that performs the best in generalizing the dataset.

4.2 ANALYSIS OF IMPACT OF COVID-19

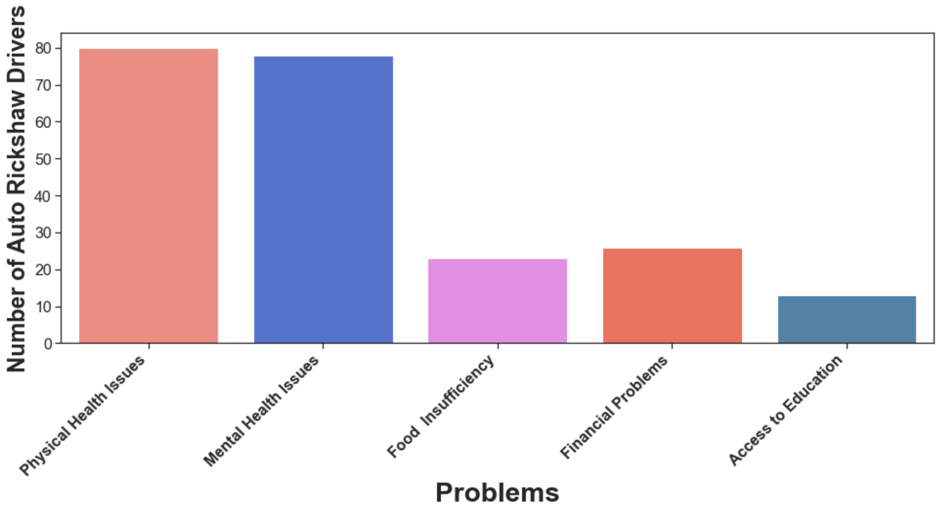
This section performs Exploratory Data Analysis on the dataset. Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. It is good to understand the data first and try to gatheras many insights from it. EDA is all about making sense of data in hand, before getting them dirty. Anything in this world could be better understood if we have an image/visualization of it. To visualize data as plots and charts to learn more about it, we used Python's data Analysis library called Pandas together with the visualization library Matplotlib. We will discuss two kinds of plots- Univariate and Multivariate. A univariate plot suggests we're only examining one variable. On the other hand, a multivariate analysis examines more than two variables. Fortwo variables, we call it bivariate.

4.2.1 Distinct Problem Elevated Due to COVID-19

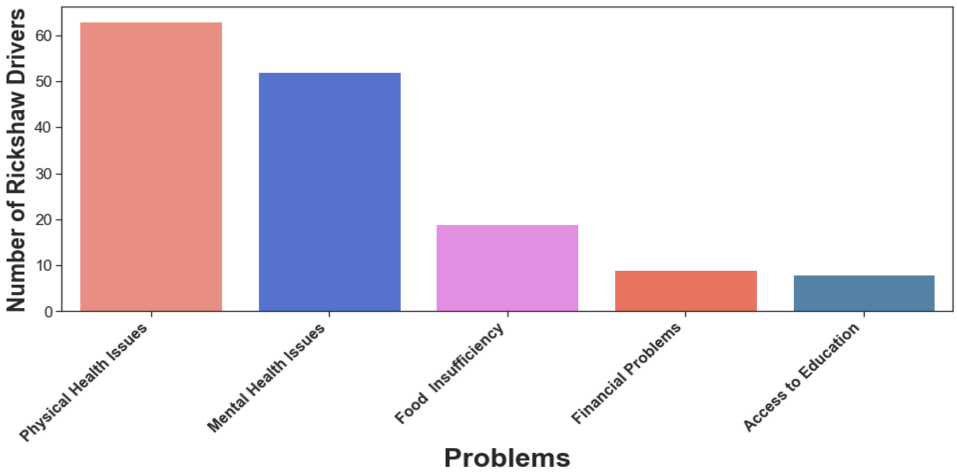
In the survey, a total of 262 CNG drivers participated, out of which 92 reported physical health issues as a difficulty, 88 reported mental health issues, 43 reported financial problems, 20 reported food insecurities, and 19 reported limited access to education as show in Figure 4.1(a). Among the 220 auto-rickshaw drivers who participated in the survey, 80 reported physical health issues, 78 reported mental health issues, 26 reported financial problems, 23 reported food insecurities, and 13 reported limited access to education show in Figure 4.1(b). It is shown in Figure 4.1(c) that a total of 151 rickshaw drivers participated in the survey, out of which 63 reported physical health issues, 52 reported mental health issues, 9 reported financial problems, 19 reported food insecurities, and 8 reported limited access to education. Among the 232 tailors who participated in the survey, 87 reported physical health issues, 66 reported mental health issues, 28 reported financial problems, 39 reported food insecurities, and 12 reported limited access to education as shown in Figure 4.1(d). In the survey, 197 construction workers participated, out of which 65 reported physical health issues, 77 reported mental health issues, 14 reported financial problems, 31 reported food insecurities, and 10 reported limited access to education as shown in Figure 4.1(e).It is shown in Figure 4.1(f) a total of 178 plumbers participated in the survey, out of which 65 reported physical health issues, 48 reported mental health issues, 41 reported financial problems, 16 reported food insecurities, and 8 reported limited access to education. Among the 210 retailers who participated in the survey, 51 reported physical health issues, 63 reported mental health issues, 67 reported financial problems, 25 reported food insecurities, and 9 reported limited access to education as shown in Figure 4.1(g). Finally, among the 215 carpenters who participated in the survey, 51 reported physical health issues, 63 reported mental health issues, 67 reported financial problems, 25 reported food insecurities, and 9 reported limited access to education in Figure 4.1(h).



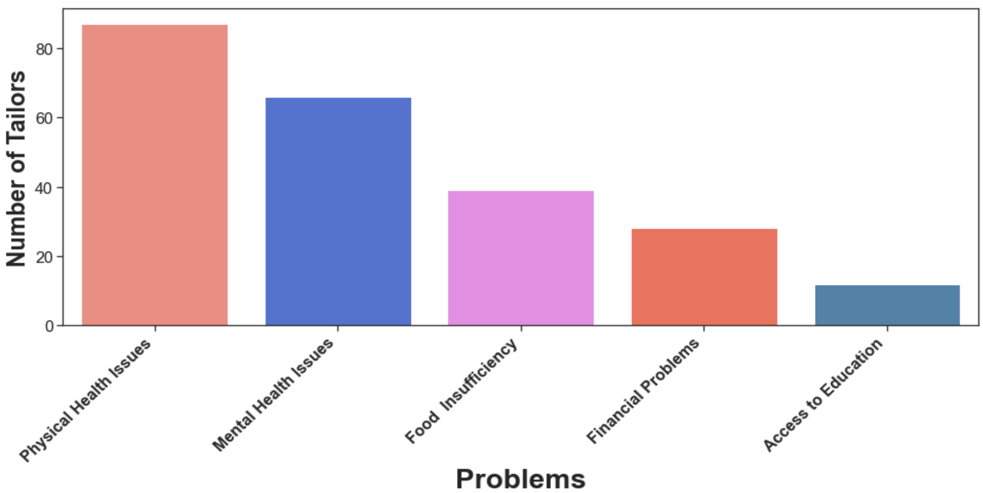
(a) Problems faced by the CNG Drivers due to COVID-19



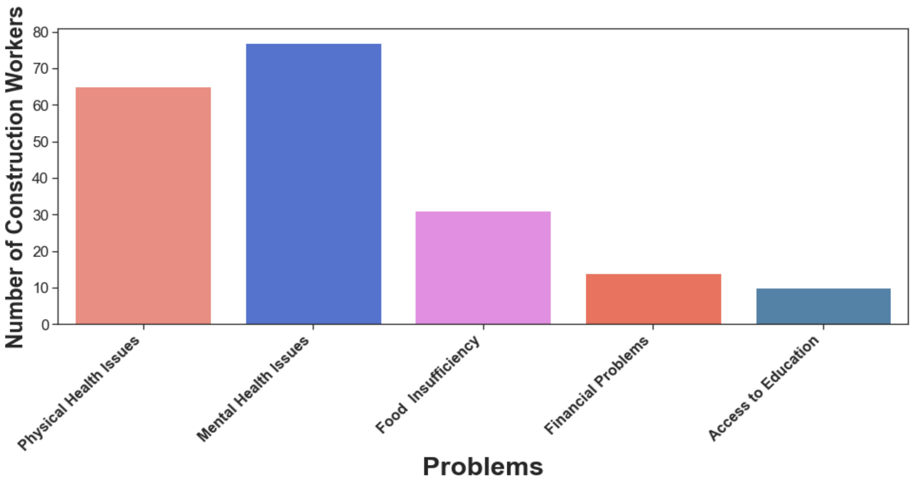
(b) Problems faced by the Auto Rickshaw Drivers due to COVID-19



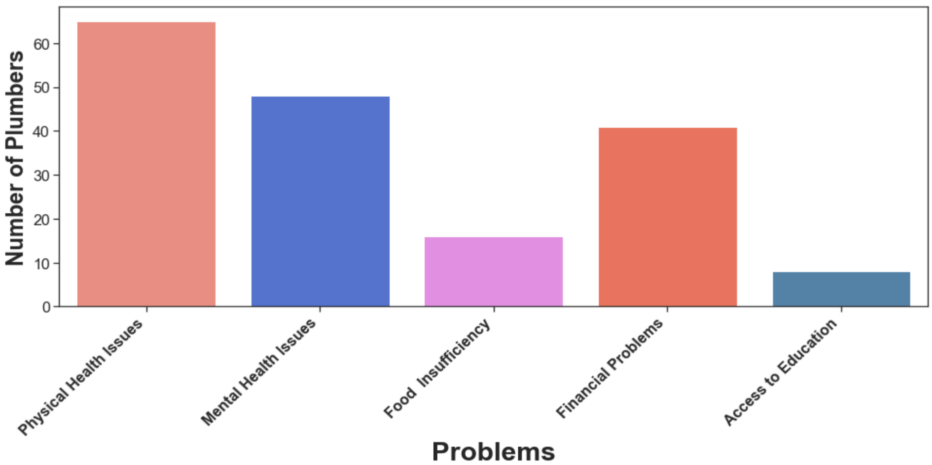
(c) Problems faced by the Rickshaw Drivers due to COVID-19



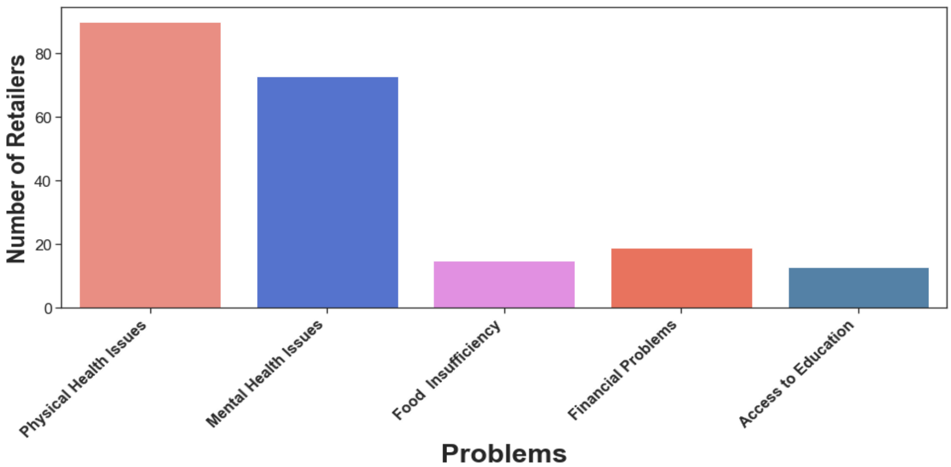
(d)Problems faced by the Tailors due to COVID-19



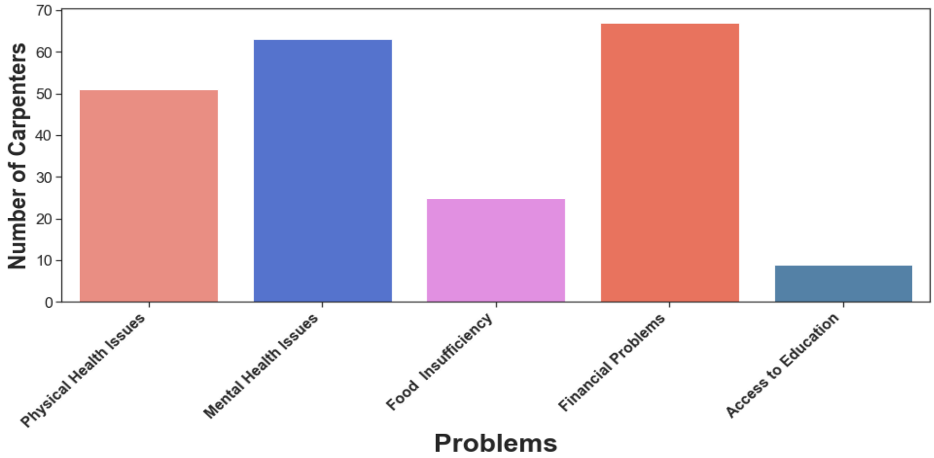
(e) Problems faced by the Construction Workers due to COVID-19



(f) Problems faced by the Plumbers due to COVID-19



(g) Problems faced by the Retailers due to COVID-19

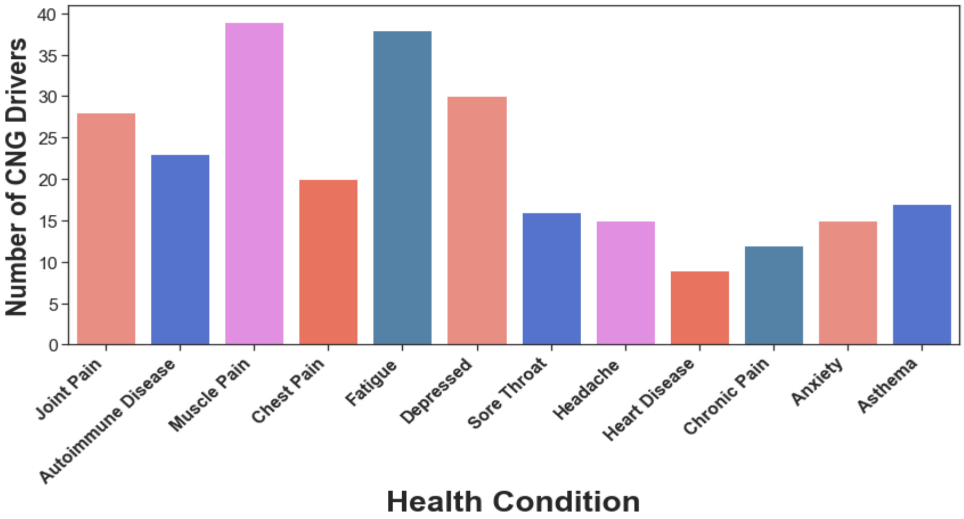


(h) Problems faced by the Carpenters due to COVID-19

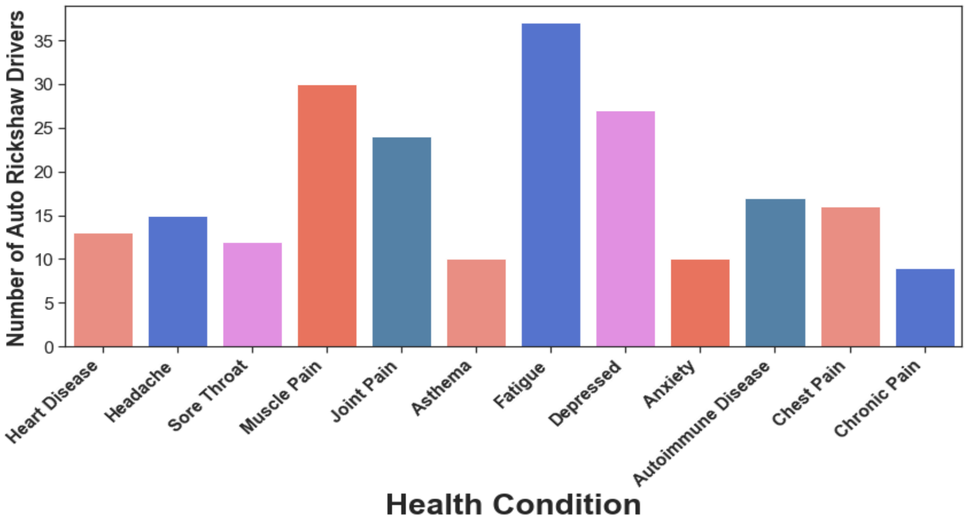
Figure 4.1: Problems raised due to COVID-19 for distinct professions.

4.2.2 Impact of COVID-19 on Health

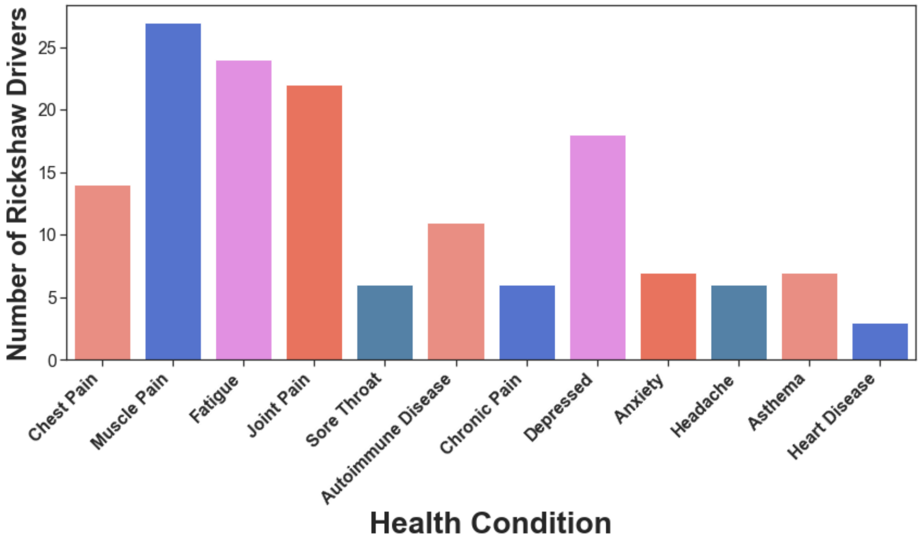
A total of 1665 daily workers participated in the survey including 232 CNG drivers, 220 autorickshaw drivers, 151 rickshaw drivers, 232 tailors, 197 construction workers, 215 carpenters, 210 retailers, and 178 plumbers. Among the CNG drivers, the highest reported conditions were muscle pain (39), fatigue (38), depression (30), and joint pain (28), followed by autoimmune disease, chest pain, asthma, sore throat, headache, anxiety, chronic pain, and heart disease shown in Figure 3.1 (a). Among autorickshaw drivers, the percentage of occurrence of fatigue, muscle pain, depression, and joint pain was higher than other health conditions shown in Figure 3.1 (b). Figure 3.1 (c) shows that among rickshaw drivers, fatigue, muscle pain, depression, and joint pain, as well as chest pain, had a higher percentage of occurrence than other health conditions. Among tailors, the highest reported conditions were muscle pain (41), fatigue (41), joint pain (33), and autoimmune disease (20), followed by chest pain, asthma, sore throat, headache, anxiety, chronic pain, and heart disease. Among construction workers, fatigue, muscle pain, depression, and joint pain, as well as headache, had a higher percentage of occurrence than other health conditions shown in Figure 3.1 (e). Figure 3.1 (f) shows that Among carpenters, fatigue and joint pain had the highest percentage of occurrence of health conditions, while heart disease had the least occurrence. Among retailers, fatigue and muscle pain had the highest percentage of occurrence of health conditions, while sore throat and chronic pain had the least occurrence shown in Figure 3.1 (g). Among plumbers, fatigue, muscle pain, depression, and joint pain had a higher percentage of occurrence of health conditions, and chronic pain had the least occurrence shown in Figure 3.1 (h).



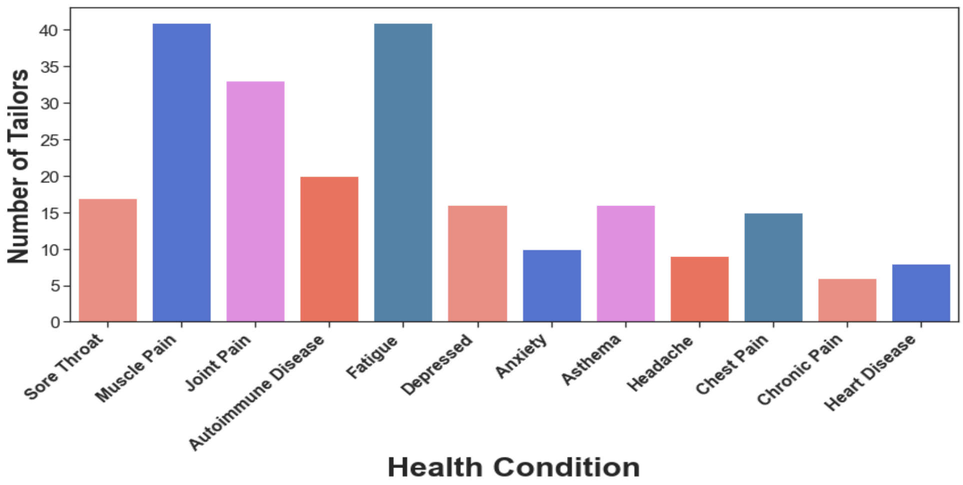
(a) Health Condition of CNG Drivers due to COVID-19



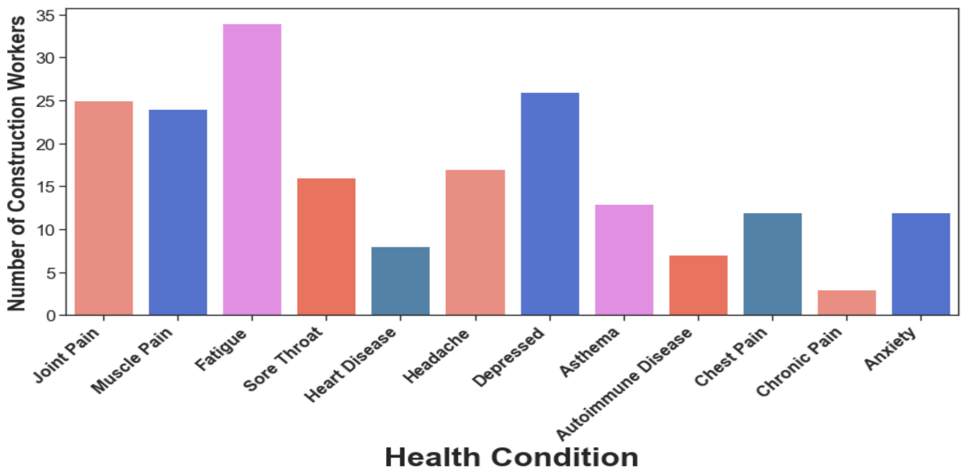
(b) Health Condition of Auto Rickshaw Drivers due to COVID-19



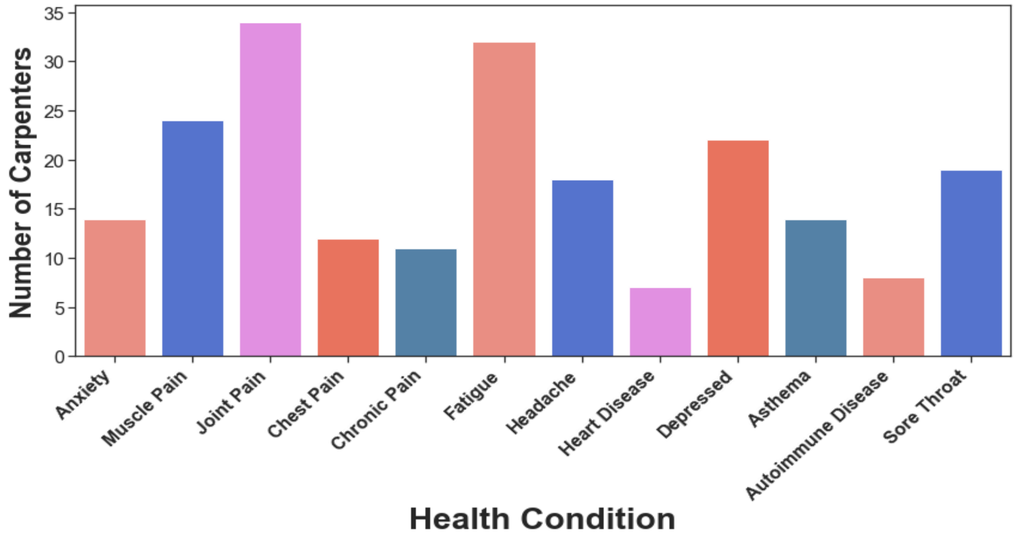
(c) Health Condition of Rickshaw Drivers due to COVID-19



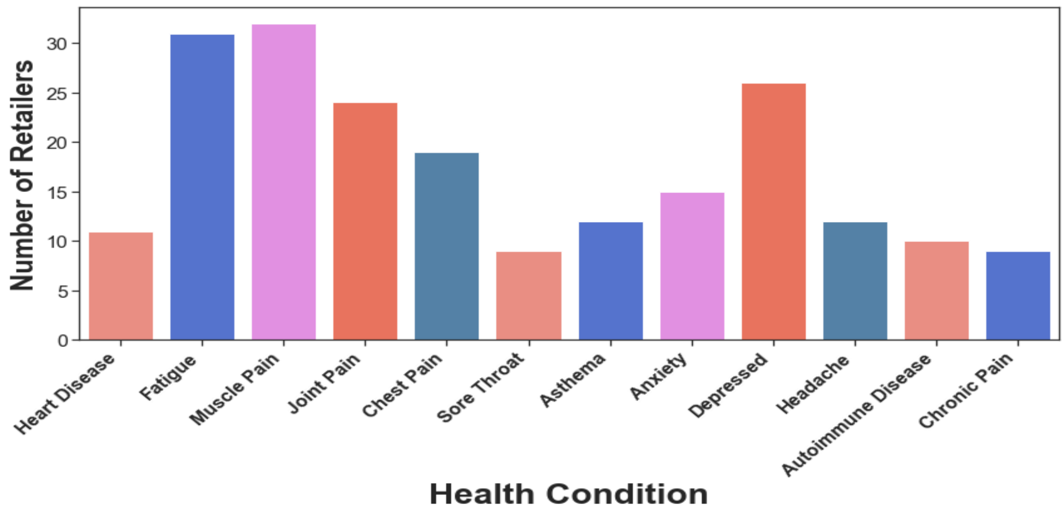
(d) Health Condition of Tailors due to COVID-19



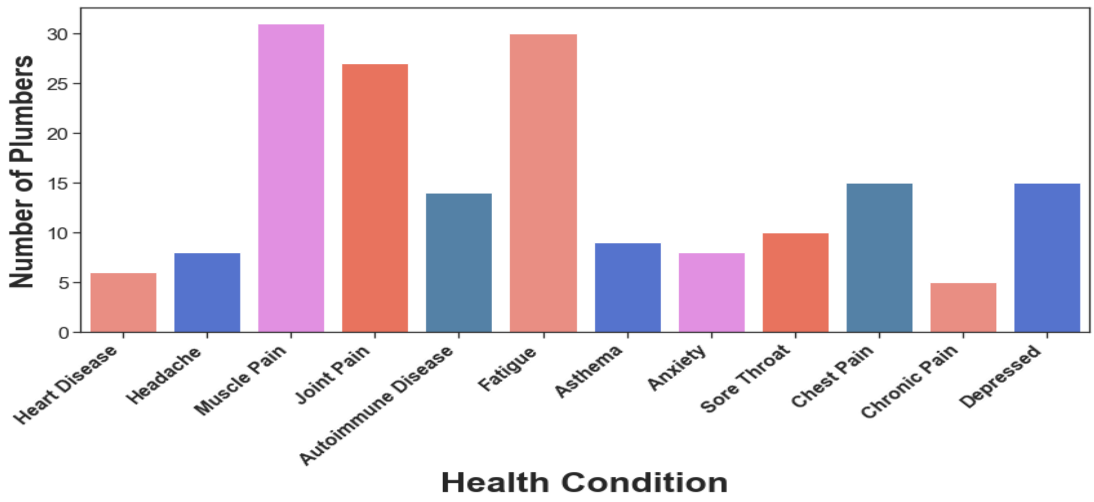
(e) Health Condition of Construction Workers due to COVID-19



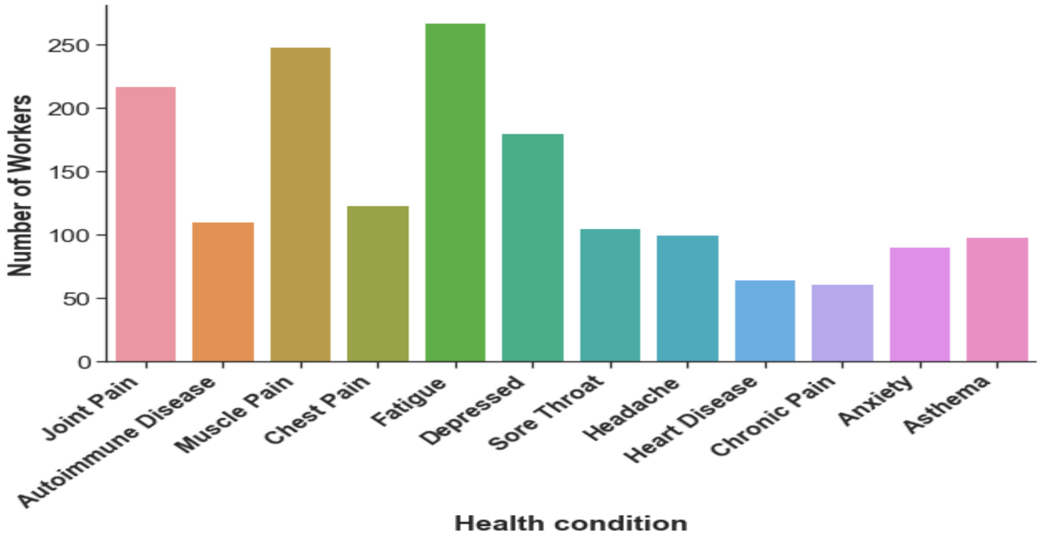
(f) Health Condition of Carpenters due to COVID-19



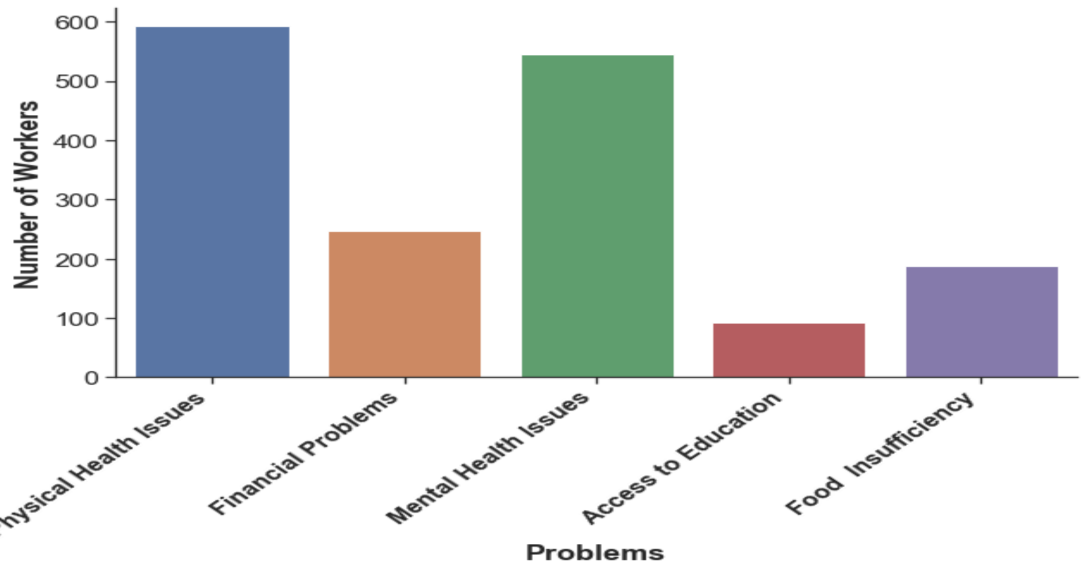
(g) Health Condition of Retailers due to COVID-19



(h) Health Condition of Plumbers due to COVID-19



(i) Physical Health Condition of Daily Worker after COVID -19

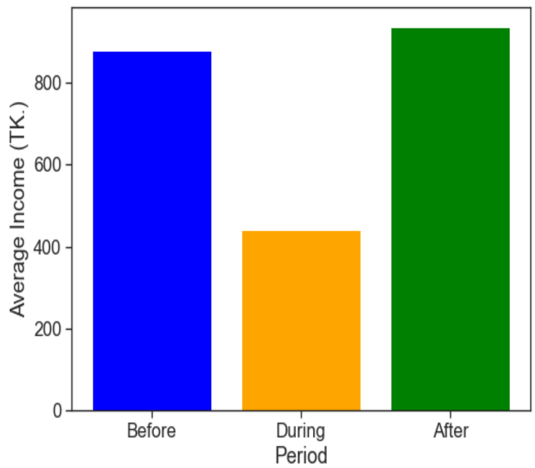


(j) Problem faced due too COVID-19

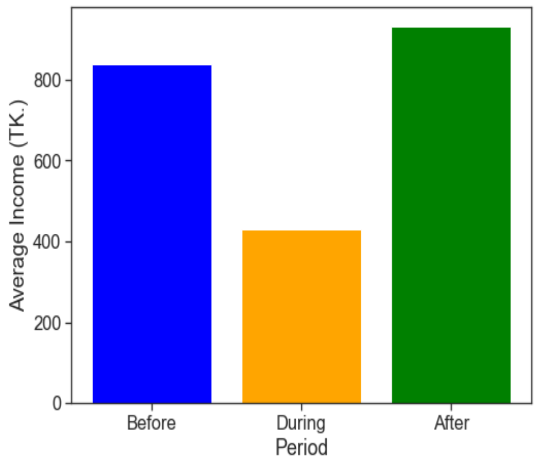
Figure 4. 2: Number of workers of different profession vs Health condition.

4.2.3 Impact of COVID-19 on Income

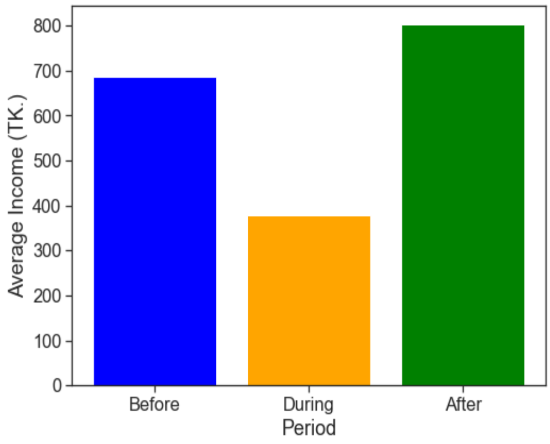
This section of the analysis presents a visual representation of the impact of COVID-19 on the average income (per day) of different types of daily wage workers over three time periods - before COVID-19, during COVID-19, and after COVID-19. For CNG drivers, it can be seen from Figure 4.3(a) that their average income decreased during COVID-19 compared to before COVID-19, and then slightly increased after COVID-19. This may be attributed to the impact of the pandemic on the transportation industry and reduced demand for CNG drivers during the pandemic due to restrictions. Similarly, it can be observed from Figure 4.3(b) that auto rickshaw drivers also experienced a decrease in average income during COVID-19, which slightly increased after COVID-19. Figure 4.3(c) shows that rickshaw drivers' average income decreased during COVID-19 but increased considerably after COVID-19. In the case of tailors, Figure 4.4(d) indicates a considerable decrease in their average income during COVID-19, which then increased considerably after COVID-19. Retailers' average income, shown in Figure 4.3(e), decreased during COVID-19 and then slightly increased after COVID-19. Carpenters, on the other hand, experienced a decrease in average income during COVID-19, which then slightly decreased after COVID-19, as seen in Figure 4.3(f). Figure 4.3(g) shows that construction workers' average income decreased during COVID-19 but then increased considerably after COVID-19. Lastly, Figure 4.3(h) indicates that auto plumbers' average income decreased during COVID-19 and remained almost the same after COVID-19. Overall, these graphs provide a visual representation of the impact of COVID-19 on the average income of daily wage workers in various occupations in Bangladesh.



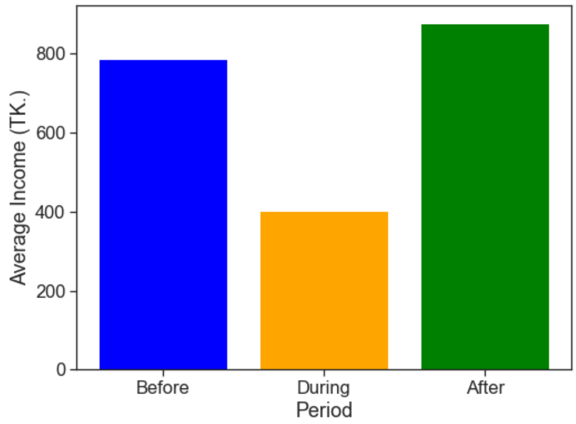
(a) Variation in Income of CNG Driver due to COVID-19



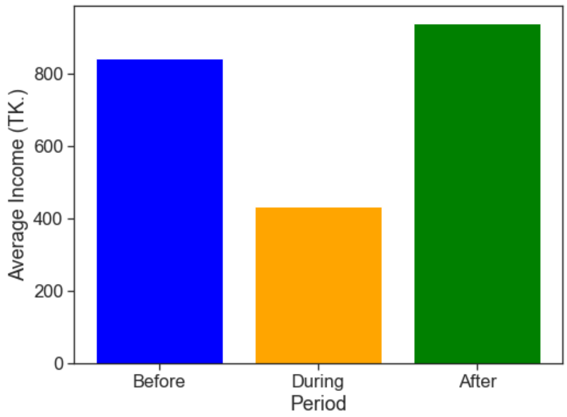
(b) Variation in Income of Auto Rickshaw Driver due to COVID-19



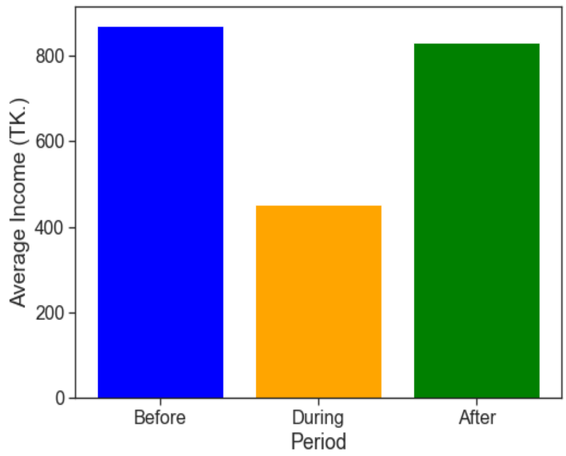
(c) Variation in Income of Rickshaw Driver due to COVID-19



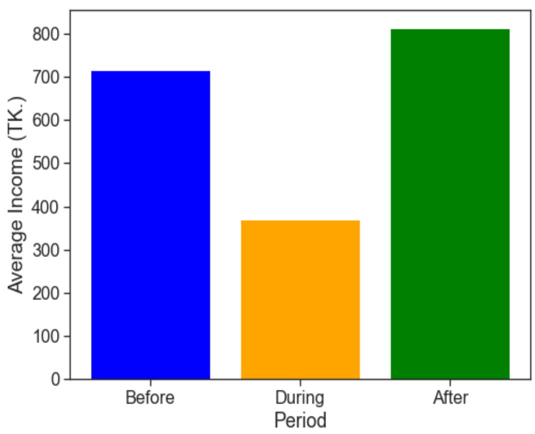
(d) Variation in Income of Tailor due to COVID-19



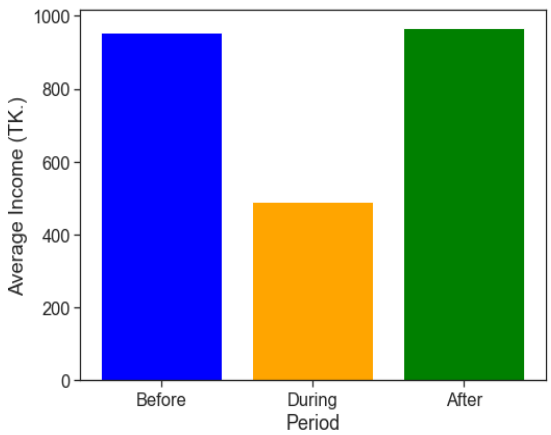
(e) Variation in Income of Retailers due to COVID-19



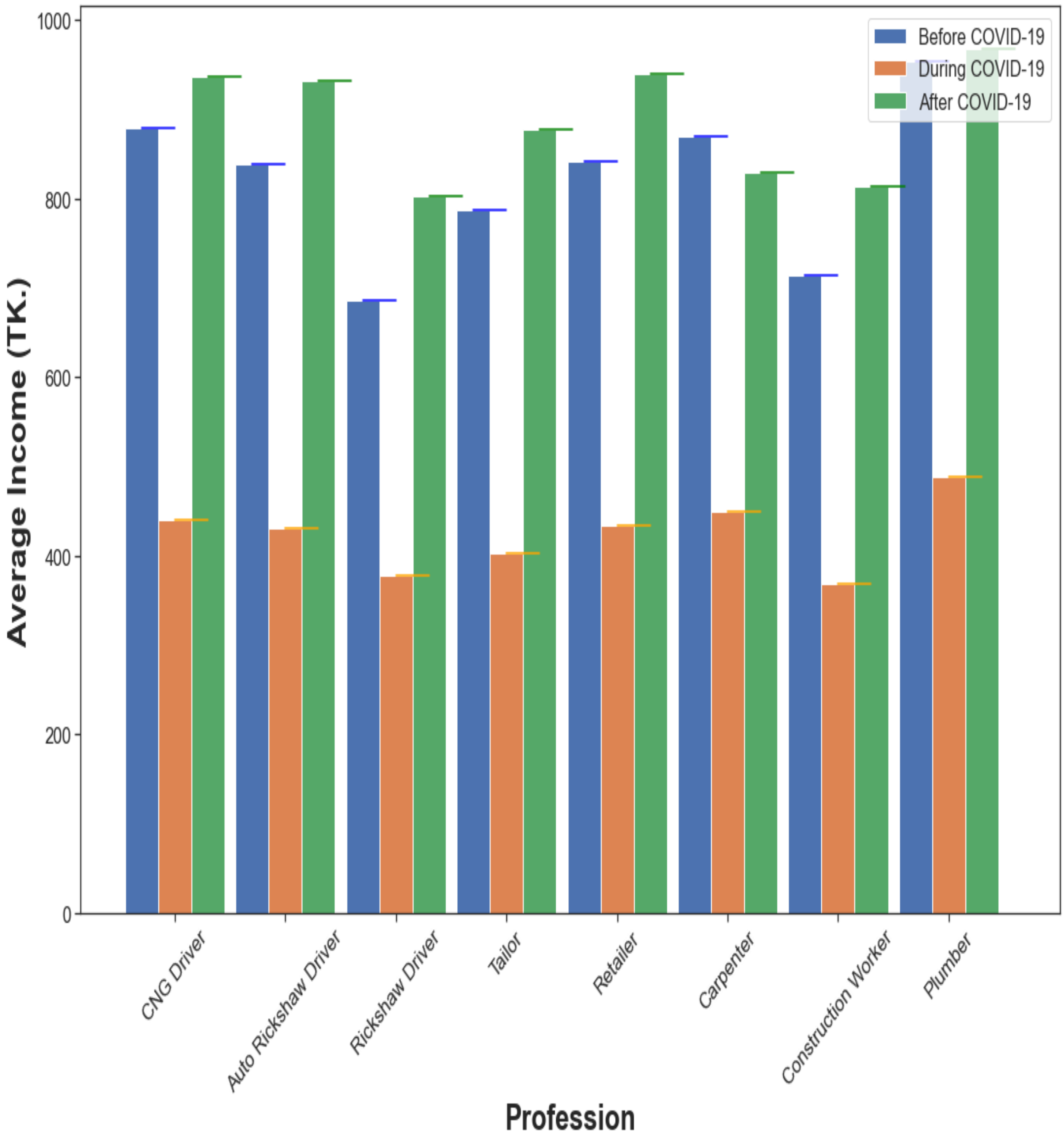
(f) Variation in Income of Carpenters due to COVID-19



(g) Variation in Income of Construction worker due to COVID-19



(h) Variation in Income of Plumber due to COVID-19

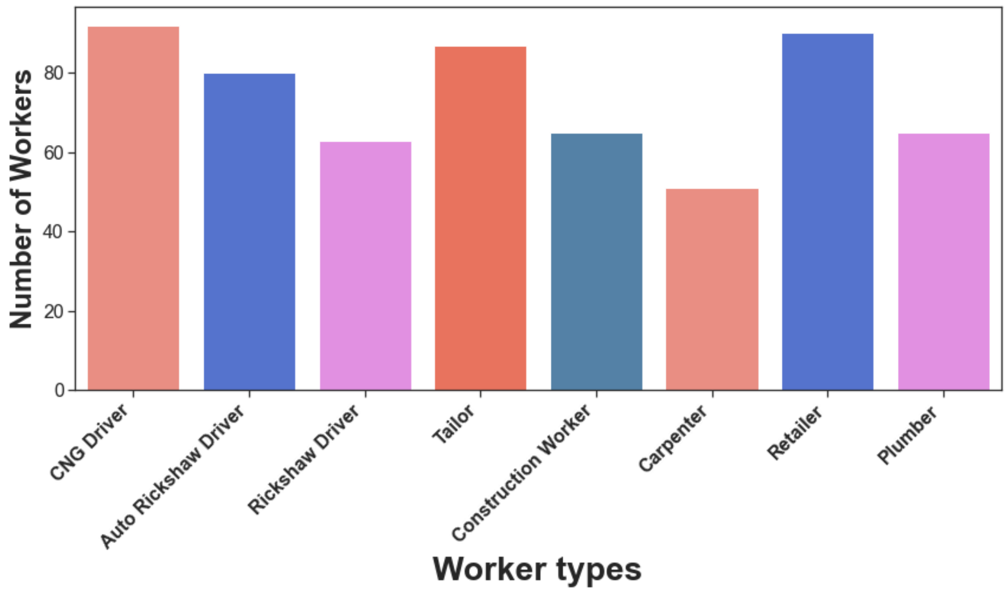


(i) Variation in income over time for all Professions.

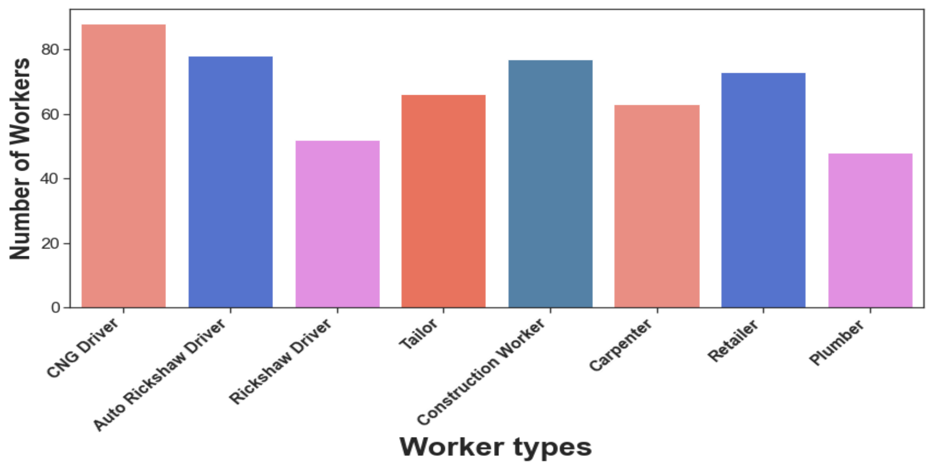
Figure 4. 3: Variation in Income due to COVID-19 for distinct professions.

4.2.4 Impact of COVID -19 for all Profession

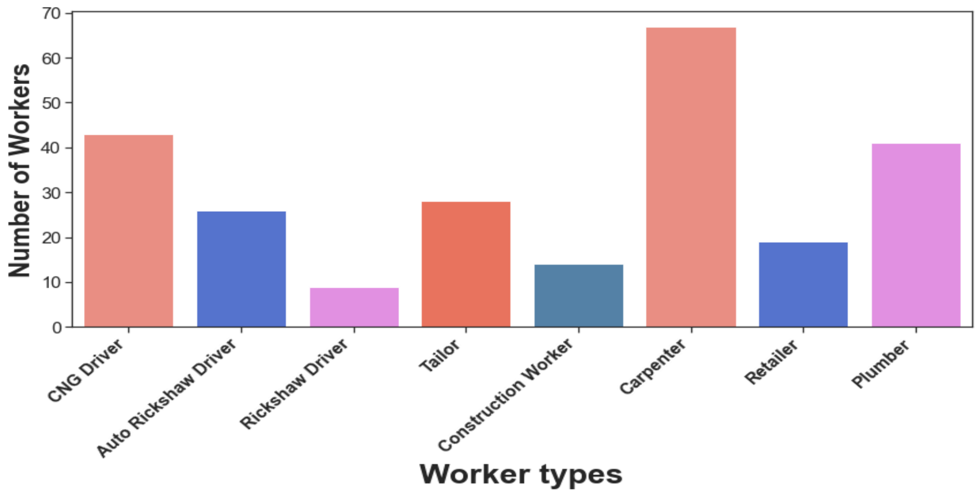
This section of the analysis focuses on the types of difficulties faced by daily wage workers due to COVID-19. Figure 4.4(a) displays the physical health issues faced by these workers. The visualization indicates that construction workers experienced the highest number of physical health issues, followed by auto-rickshaw drivers, rickshaw drivers, and tailors. This could be due to the nature of their work, which may involve greater exposure to COVID-19, and a lack of access to personal protective equipment (PPE) or medical facilities. Figure 4.4(b) presents the number of daily workers who reported facing mental health issues due to the COVID-19 pandemic. The graph shows that construction workers and auto-rickshaw drivers reported the highest number of mental health issues, while retailers and carpenters reported the lowest. Figure 4.4(c) illustrates the number of daily workers in each profession who reported facing financial problems due to COVID-19. Among the daily workers surveyed, the retailer profession had the highest number of workers (67) reporting financial problems, followed by the plumber profession (41) and the construction worker profession (14). Figure 4.4(d) shows the number of daily workers facing food insecurities due to the COVID-19 pandemic. The graph highlights that daily workers in the construction, transportation, and service sectors are the most affected by food insecurities due to the COVID-19 pandemic. These workers are likely to be in lower-paying jobs, which makes it difficult for them to access adequate food during the pandemic. Figure 4.4(e) indicates the number of daily workers from different professions who faced limited access to education due to COVID-19. The graph shows that daily workers in the transportation and service sectors are the most affected by limited access to education due to the COVID-19 pandemic.



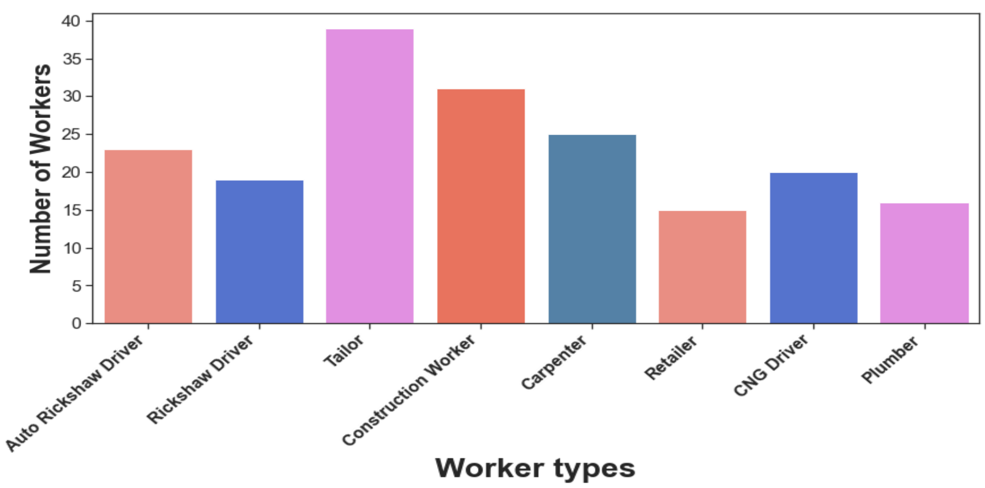
(a) Physical Health Issues faced by Daily Workers due to COVID-19



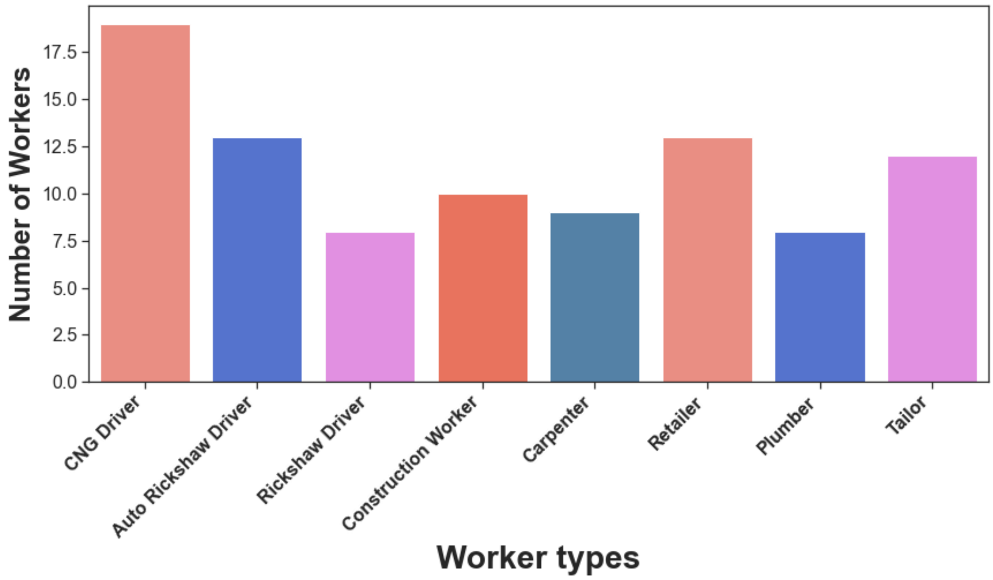
(b) Mental Health Issues faced by Daily Workers due to COVID-19



(c) Financial Problem faced by Daily Workers due to COVID-19



(d) Food Insufficiency faced by Daily Workers due to COVID-19



(e) Access to Education faced by Daily Workers due to COVID-19

Figure 4. 4:(a) Physical health issues vs Profession, (b) Mental health issues vs Profession ,(c) Financial Problems vs Profession, (d) Food Insufficiency vs Profession, (e Access to education vs Profession.

4.3 EFFECT OF COVID-19 PREDICTION USING SEVERAL ALGORITHMS

In this paper, we employed seven different machine learning algorithms to construct a predictive model. The algorithms used were logistic regression, decision tree, random forest, support vector machine (SVM), AdaBoost, k-nearest neighbors and XGBoost. We evaluated the performance of each model, and the results are presented below:

4.3.1 Logistic Regression

a) Performance Metrics of Logistic Regression

Accuracy: 0.591

Table 4. 1 Performance Metrics of Logistic Regression

	Precision Value	Recall	F1 Score	Support
0	0.00	0.00	0.00	45
1	0.69	0.69	0.69	37
2	0.84	0.73	0.78	24
3	0.55	0.59	0.57	116
4	0.55	0.63	0.59	111
Accuracy			0.59	333
Macro Avg	0.53	0.53	0.53	333
Weighted Avg	0.56	0.59	0.57	333

b) Confusion Matrix:

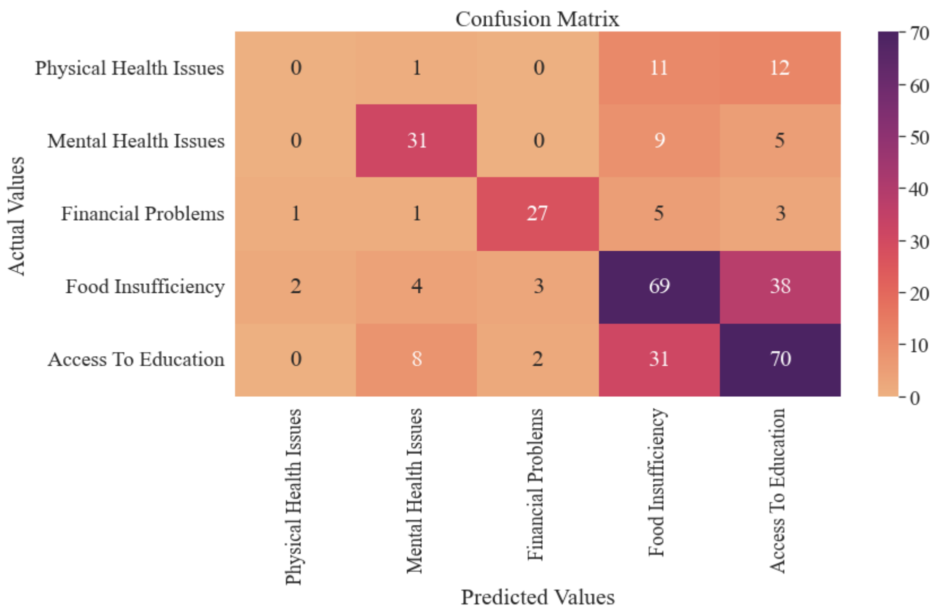


Figure 4. 5: Confusion Matrix of Logistic Regression

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled "Access to Education" and "Food Insufficiency". The class labeled "Financial Problems", "Mental Health Issues" are classified correctly. The model doesn't seem to generalize well in the case of these two classes. The class labeled " Physical Health Issues" is not classified correctly.

4.3.2 Random Forest

a) Performance Metrics of Random Forest

Accuracy: 0.924

Table 4. 2: Performance Metrics of Random Forest

	Precision Value	Recall	F1 Score	Support
0	0.95	0.79	0.86	24
1	0.95	0.87	0.91	37
2	0.91	0.81	0.86	37
3	0.92	0.97	0.95	116
4	0.92	0.96	0.94	111
Accuracy			0.92	333
Macro Avg	0.93	0.88	0.90	333
Weighted Avg	0.93	0.92	0.92	333

(b)Confusion Matrix:

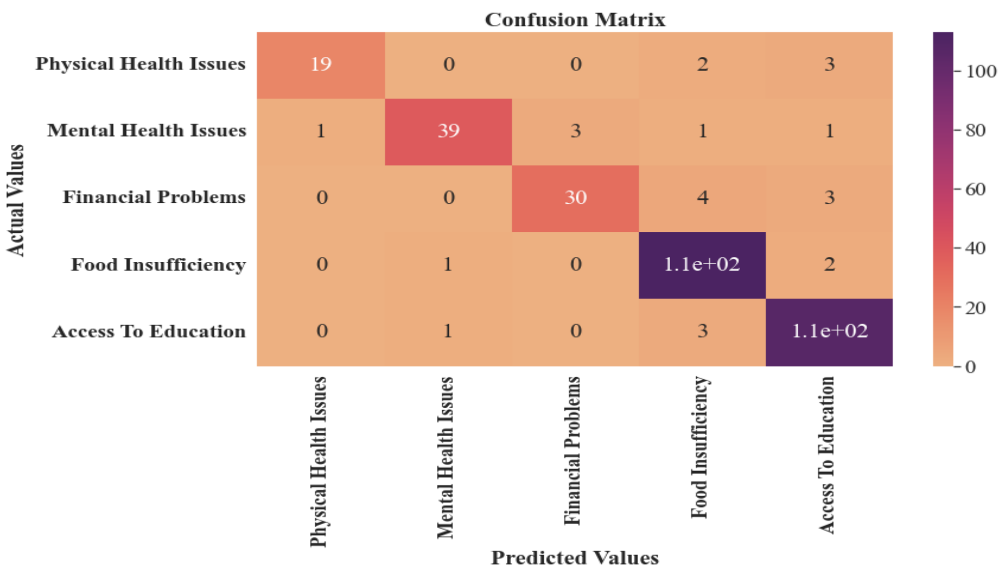


Figure 4. 6: Confusion Matrix of Random Forest Model.

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled "Access To Education" and

"Food Insecurities". The class labeled "Financial Problems", "Mental Health Issues" and "Physical Health Issues" are classified correctly .The model doesn't seem to generalize well in the case of these three classes.

4.3.3 Decision Tree

a) Performance Metrics of Decision Tree

Accuracy 0.906

Table 4. 3: Performance Metrics of Decision Tree

	Precision Value	Recall	F1 Score	Support
0	0.89	0.71	0.79	24
1	0.89	0.76	0.82	45
2	0.84	0.84	0.84	37
3	0.93	0.96	0.94	116
4	0.91	0.97	0.94	111
Accuracy			0.90	333
Macro avg	0.88	0.85	0.87	333
Weighted avg	0.90	0.90	0.90	333

b) Confusion Matrix:

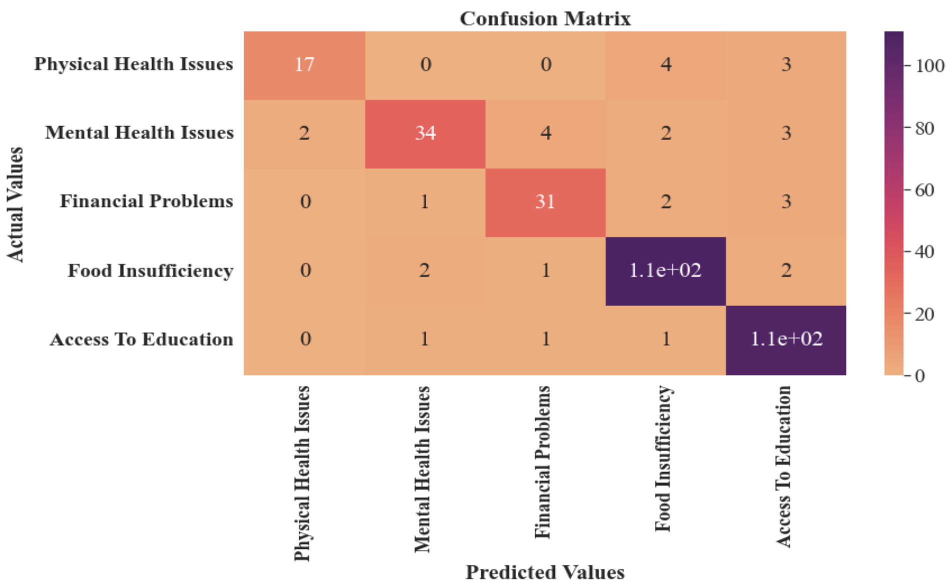


Figure 4. 7: Confusion Matrix of Decision Tree Model.

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled “Access To Education” and “Food Insufficiency”. The class labeled “Physical Health Issues”, “Mental Health Issues” and “Financial Problems” are classified correctly .The model doesn’t seem to generalize well in the case of these three classes.

4.3.4 Support Vector Machine

a) Performance Metrics of Support Vector Machine

Accuracy: 0.72

Table 4. 4: Performance Metrics of Support Vector Machine**.**

	Precision Value	Recall	F1 Score	Support
0	1.00	0.04	0.08	24
1	0.82	0.73	0.78	45
2	0.96	0.73	0.83	37
3	0.64	0.89	0.75	116
4	0.73	0.68	0.71	111
Accuracy			0.72	333
Macro Avg	0.83	0.62	0.63	333
Weighted Avg	0.76	0.72	0.70	333

b) Confusion Matrix:

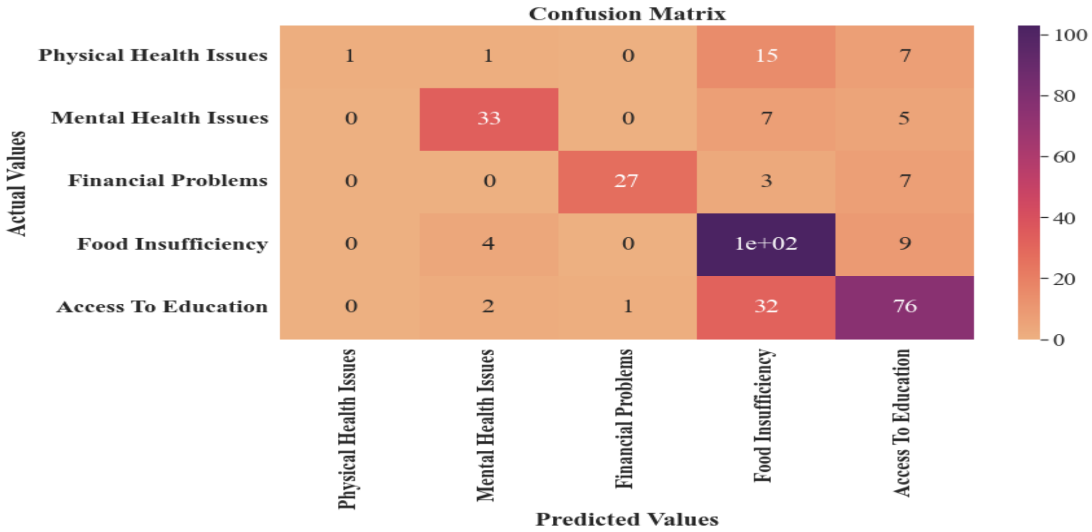


Figure 4. 7: Confusion Matrix of Support Vector Machine.

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled “Access To Education” and “Food Insufficiency”. The class labeled “Financial Problems”, “Mental Health Issues” are classified correctly .The model doesn’t seem to generalize well in the case of these two classes. The class labeled “Physical Health Issues” is not classified correctly.

4.3.5 K-Nearest Neighbors

a) Performance Metrics of K-NN

Accuracy: 0.636

Table 4. 5: Performance Metrics of KNN .

	Precision Value	Recall	F1 Score	Support
0	0.42	0.21	0.28	24
1	0.71	0.78	0.74	45
2	0.83	0.68	0.75	37
3	0.58	0.72	0.64	116
4	0.64	0.58	0.61	111
Accuracy			0.64	333
Macro Avg	0.64	0.59	0.60	333
Weighted Avg	0.64	0.64	0.63	333

b) Confusion Matrix:

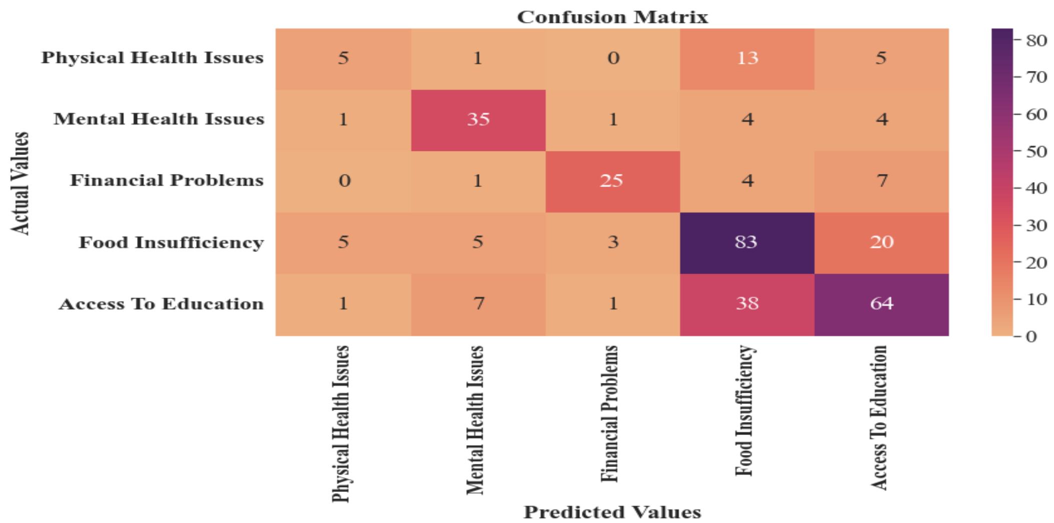


Figure 4. 8: Confusion Matrix of KNN Model.

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled “Limited Access To Education” and “Food Insecurities”. The class labeled “Financial Problems”, “Mental Health Issues” are classified correctly .The model doesn’t seem to generalize well in the case of these two classes. The class labeled “Physical Health Issues” is not classified correctly.

4.3.6 AdaBoost

a) Performance Metrics of AdaBoost

Accuracy: 0.648

Table 4. 6: Performance Metrics of AdaBoost

	Precision Value	Recall	F1-Score	Support
0	0.25	0.17	0.20	24
1	0.51	0.53	0.52	45
2	0.83	0.14	0.23	37
3	0.70	0.76	0.73	116
4	0.69	0.86	0.76	114
Accuracy			0.65	333
Macro Avg	0.60	0.49	0.49	333
Weighted Avg	0.65	0.65	0.62	333

(b) Confusion Matrix

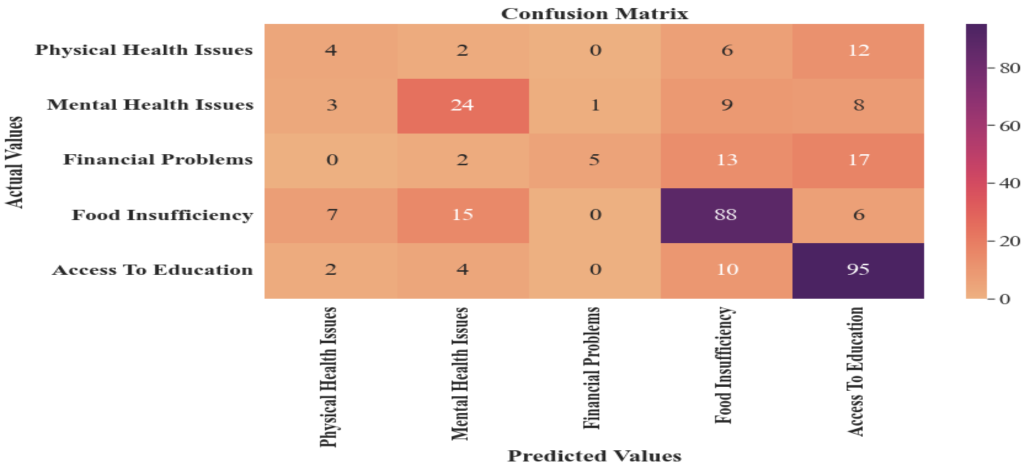


Figure 4. 9: Confusion Matrix of Ada Boost .

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled “Access To Education” and “Food Insufficiency”. The class labeled “Mental Health Issues” is classified correctly .The model doesn’t seem to generalize well in the case of this class .The class labeled “Financial Problems” and “Physical Health Issues” are not classified correctly.

4.3.7 XGBoost

a) Performance Metrics of XGBoost

Accuracy: 0.912

Table 4. 7: Performance Metrics of XGboost

	Precision Value	**Recall**	**F1-Score**	**Support**
0	1.00	0.62	0.77	24
1	0.86	0.82	0.84	45
2	0.94	0.81	0.87	37
3	0.93	0.97	0.95	116
4	0.90	0.98	0.94	111
Accuracy			0.91	333
Macro Avg	0.93	0.84	0.87	333
Weighted Avg	0.92	0.91	0.91	333

b) Confusion Matrix

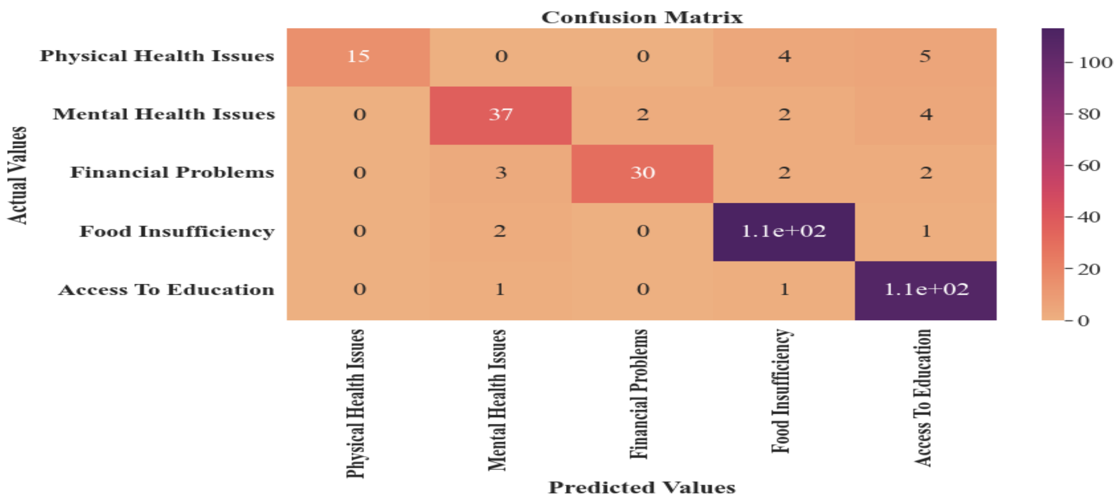


Figure 4. 10: Confusion Matrix of XGBoost.

Figure shows the confusion matrix to check how well our classification models are performing and what kinds of errors are they making. From the figure, we observe that the algorithm performs great in classifying the classes labeled “Limited Access To Education” and “Food Insecurities”. The class labeled “Physical Health Issues”, “Mental Health Issues” and “Financial Problems” are classified correctly .The model doesn’t seem to generalize well in the case of these three classes.

4.4. COMPARISON OF PERFORMANCE METRICES AMONG DISTINCT PREDICTION ALGORITHMS

To determine the best algorithm for classifying the difficulties faced due to COVID-19, the aim of this research was to conduct a comparative analysis of different algorithms.

Table 4. 8: Performance Metrices of Different Algorithm

Algorithm	Accuracy	Precision Value	Recall	F1-Score
Logistic Regression	59.1%	56.1%	59.1%	57.4%
Random Forest	**92.4%**	**92.5%**	**92.4%**	**92.3%**
Decision Tree	90.3%	90.3%	90.3%	90.1%
Support Vector Machine	72%	75.8%	72%	69.8%
KNN	63.6%	63.6%	63.6%	62.9%
Ada Boost	64.8%	65.2%	64.8%	61.8%
XGBoost	91.2%	91.5%	91.2%	90.9%

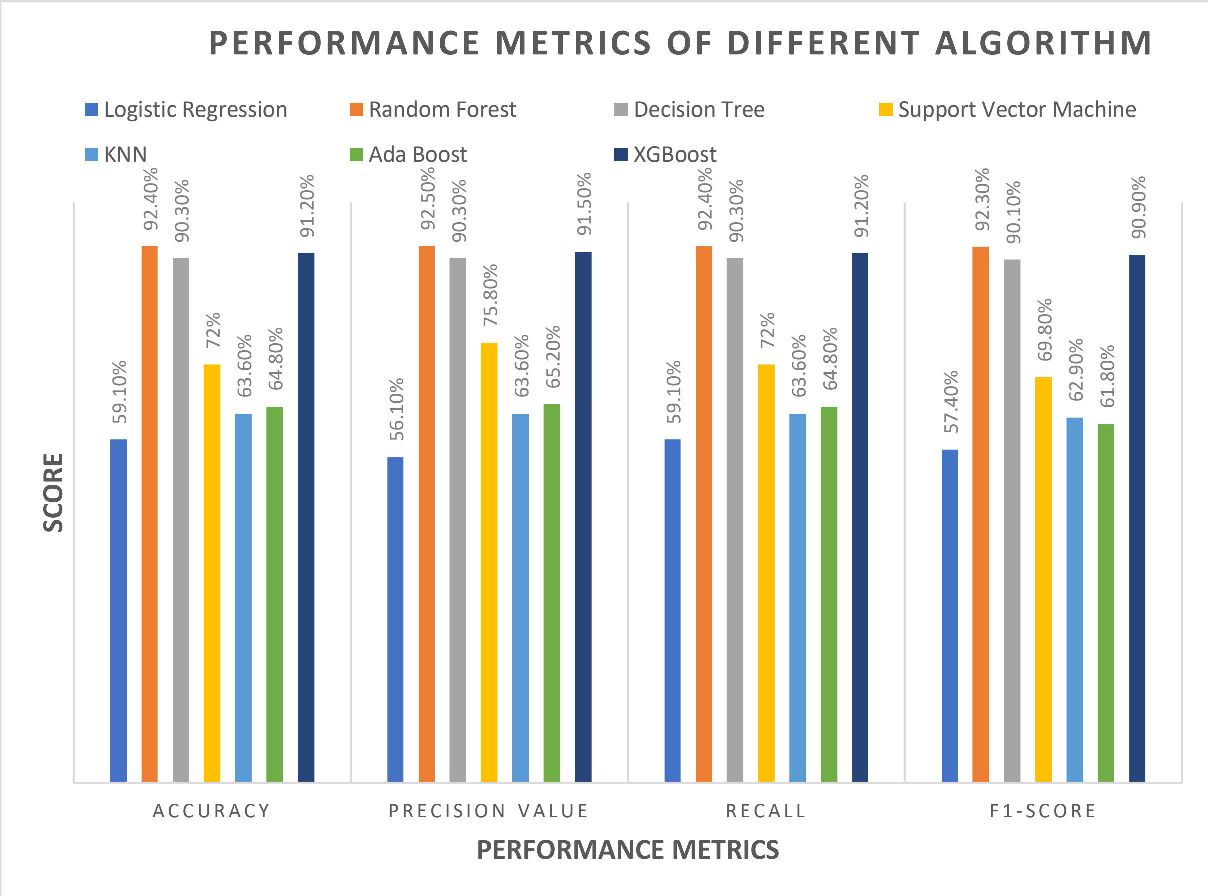


Figure 4. 11: Performance Metrices of Different Algorithm

4.5 FINAL PREDICTION MODEL

This study utilized seven machine learning algorithms, namely Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, K-NN, AdaBoost, and XGBoost, to predict the impact of COVID-19 on the daily wage workers. Among them Random Forest achieved the highest accuracy rates of 92% , also observed that XGBoost, and Decision Tree performed the best in terms of accuracy. However, Support Vector Machine, K-NN, and Ada Boost also showed promising results. On the other hand, Logistic Regression did not perform well on our dataset. The effectiveness of an algorithm is determined by various parameters, and selecting the appropriate algorithm for a particular dataset is crucial in obtaining accurate predictions. Our findings suggest that careful selection of classification algorithms is necessary for accurately predicting the impacts of COVID-19 on the lives of daily wage workers using machine learning algorithms.

CHAPTER 5

CONCLUSION

5.1 INTRODUCTION

The findings of this research highlight the significant impact of the COVID-19 pandemic on daily wage workers, particularly those in the construction, transportation, and service sectors. The research also shows that different professions experienced different types and degrees of impact from the pandemic. The study also provides insights into the effectiveness of different machine learning algorithms in predicting the impact of COVID-19 on daily wage workers.

5.2 CONCLUSION

This study conducted to predict the impact of COVID-19 on daily wage workers in the construction, transportation, and service sectors, and to analyze the impact of COVID-19 on factors such as health condition and income:

- Seven machine learning algorithms were utilized, including Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, K-NN, AdaBoost, and XGBoost, to accurately predict the impacts of COVID-19 on daily wage workers.
- Random Forest and XGBoost models achieved the highest accuracy rates of 92% and 91%, respectively.
- The study analyzed the impact of COVID-19 on several factors including health condition, income before, during, and after COVID-19, and the problems faced by daily wage workers.
- The research provides valuable insights into how these factors have affected daily wage workers in the construction, transportation, and service sectors, which can assist policymakers in developing targeted strategies to address the challenges faced by this population.

- The methodology and insights gained from this research can be applied to other sectors and populations affected by the pandemic, which can result in more informed decision-making and better outcomes.
- The utilization of machine learning algorithms enables precise and dependable predictions of the impact of COVID-19 on daily wage workers.

In conclusion, this study highlights the significance of utilizing machine learning algorithms to comprehend the impact of COVID-19 on vulnerable populations and design effective policies to support them.

5.3 FUTURE SCOPE OF THE STUDY

As we have used a small dataset, the model's accuracy can be further enhanced by using a large amount of data for training purposes. Other robust or hybrid algorithms (a combination of two or more algorithms) can be implemented for increasing accuracy.

REFERENCES

[1] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 700–704. doi: 10.1109/ICCS45141.2019.9065812.

[2] M. Alrahhal and K. P. Supreethi, "Content-Based Image Retrieval using Local Patterns and Supervised Machine Learning Techniques," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Feb. 2019, pp. 118–124. doi: 10.1109/AICAI.2019.8701255.

[3] S. Subramanian, B. Tseng, R. Barbieri, and E. N. Brown, "Unsupervised Machine Learning Methods for Artifact Removal in Electrodermal Activity," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 399–402. doi: 10.1109/EMBC46164.2021.9630535.

[4] G. Srikanth, N. Nukavarapu, and S. Durbha, "Deep Reinforcement Learning Interdependent Healthcare Critical Infrastructure Simulation model for Dynamically Varying COVID-19 scenario - A case study of a Metro City," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 8499–8502. doi: 10.1109/IGARSS47720.2021.9554875.

[5] K. Liu *et al.*, "Using mobile phone big data to discover the spatial patterns of rural migrant workers' return to work in China's three urban agglomerations in the post-COVID-19 era," *Environ Plan B Urban Anal City Sci*, p. 239980832110693, Feb. 2022, doi: 10.1177/23998083211069375.

[6] N. Barker *et al.*, *Migration and the labour market impacts of COVID-19*, vol. 2020. UNU-WIDER, 2020. doi: 10.35188/UNU-WIDER/2020/896-2.

[7] J. Maljaars *et al.*, "Impact of the COVID-19 Pandemic on Daily Life: Diverse Experiences for Autistic Adults," *J Autism Dev Disord*, vol. 53, no. 1, pp. 319–331, Jan. 2023, doi: 10.1007/s10803-022-05437-x.

[8] F. Research, S. Bourgault, A. Peterman, and M. O'donnell, "Center for Global Development Violence Against Women and Children During COVID-19-One Year On and 100 Papers In A."

[9] D. Štreimikienė, T. Baležentis, A. Volkov, E. Ribašauskienė, M. Morkūnas, and A. Žičkienė, "Negative effects of covid-19 pandemic on agriculture: systematic

literature review in the frameworks of vulnerability, resilience and risks involved," *Economic Research-Ekonomska Istraživanja*, vol. 35, no. 1, pp. 529–545, Dec. 2022, doi: 10.1080/1331677X.2021.1919542.

[10] Md. H. al Banna *et al.*, "The impact of the COVID-19 pandemic on the mental health of the adult population in Bangladesh: a nationwide cross-sectional study," *Int J Environ Health Res*, vol. 32, no. 4, pp. 850–861, Apr. 2022, doi: 10.1080/09603123.2020.1802409.

[11] S. K. Deb and S. Md. Nafi, "Impact of COVID-19 Pandemic on Tourism: Perceptions from Bangladesh," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3632798.

[12] S. Cheval, C. Mihai Adamescu, T. Georgiadis, M. Herrnegger, A. Piticar, and D. R. Legates, "Observed and Potential Impacts of the COVID-19 Pandemic on the Environment," *Int J Environ Res Public Health*, vol. 17, no. 11, p. 4140, Jun. 2020, doi: 10.3390/ijerph17114140.

[13] G. McCartney, C. L. Chi In, and J. S. de A. F. Pinto, "COVID-19 impact on hospitality retail employees' turnover intentions," *International Journal of Contemporary Hospitality Management*, vol. 34, no. 6, pp. 2092–2112, May 2022, doi: 10.1108/IJCHM-08-2021-1053.

[14] G. M. M. Alam and M. N. Khatun, "Impact of COVID-19 on vegetable supply chain and food security: Empirical evidence from Bangladesh," *PLoS One*, vol. 16, no. 3, p. e0248120, Mar. 2021, doi: 10.1371/journal.pone.0248120.

[15] M. Mhango, M. Dzobo, I. Chitungo, and T. Dzinamarira, "COVID-19 Risk Factors Among Health Workers: A Rapid Review," *Saf Health Work*, vol. 11, no. 3, pp. 262–265, Sep. 2020, doi: 10.1016/j.shaw.2020.06.001.

[16] H. Shen, M. Fu, H. Pan, Z. Yu, and Y. Chen, "The Impact of the COVID-19 Pandemic on Firm Performance," *Emerging Markets Finance and Trade*, vol. 56, no. 10, pp. 2213–2230, Aug. 2020, doi: 10.1080/1540496X.2020.1785863.

[17] H. Kabir, M. Maple, and K. Usher, "The impact of COVID-19 on Bangladeshi readymade garment (RMG) workers," *J Public Health (Bangkok)*, vol. 43, no. 1, pp. 47–52, Apr. 2021, doi: 10.1093/pubmed/fdaa126.

[18] A. Kumar, S. Luthra, S. K. Mangla, and Y. Kazançoğlu, "COVID-19 impact on sustainable production and operations management," *Sustainable Operations and Computers*, vol. 1, pp. 1–7, 2020, doi: 10.1016/j.susoc.2020.06.001.

[19] A. R. Masonbrink and E. Hurley, "Advocating for Children During the COVID-19 School Closures," *Pediatrics*, vol. 146, no. 3, Sep. 2020, doi: 10.1542/peds.2020-1440.

[20] L. Sher, "The impact of the COVID-19 pandemic on suicide rates," *QJM: An International Journal of Medicine*, vol. 113, no. 10, pp. 707–712, Oct. 2020, doi: 10.1093/qjmed/hcaa202.

[21] S. F. Ardabili *et al.*, "COVID-19 Outbreak Prediction with Machine Learning," *Algorithms*, vol. 13, no. 10, p. 249, Oct. 2020, doi: 10.3390/a13100249.

[22] S. I. Ao, International Association of Engineers, WCECS (2008.10.22-24 San Francisco), and World Congress on Engineering and Computer Science (2008.10.22-24 San Francisco), *WCECS 2008, World Congress on Engineering and Computer Science, San Francisco, USA, 22-24 October, 2008*. IAENG International Association of Engineers, 2008.

[23] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[24] "Decision Tree Algorithm in Machine Learning - Javatpoint." https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed Mar. 04, 2023).

[25] "Support Vector Machine (SVM) Algorithm - Javatpoint." https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (accessed Mar. 04, 2023).

[26] "K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint." https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Mar. 04, 2023).

[27] B. Kégl, "The return of ADABOOST.MH: Multi-class Hamming trees," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[28] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results Phys*, vol. 27, p. 104462, Aug. 2021, doi: 10.1016/j.rinp.2021.104462.